Spatial dissection of a soundfield using spherical harmonic decomposition

Abdullah Fahim

October 2019

A thesis submitted for the degree of Doctor of Philosophy
of The Australian National University

Research School of Electrical, Energy and Materials Engineering
College of Engineering and Computer Science
The Australian National University

To my beloved parents, Shahid and Nahar,
and my loving wife, Samiya.

Declaration

The contents of this thesis are the results of original research and have not been submitted for a higher degree to any other university or institution. Much of this work has either been published as journal/conference papers or submitted for publications as journal papers. Following is a list of these papers:

Journal Articles

- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "PSD Estimation and Source Separation in a Noisy Reverberant Environment using a Spherical Microphone Array", *IEEE/ACM Transactions on Audio Speech and Language Processing*, Volume 26, Issue 9, pp. 1594–1607, 2018.
- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "Single-Channel Speech Dereverberation in Noisy Environment for Non-Orthogonal Signals". *Acta Acustica united with Acustica*, Volume 104, Issue 6, pp. 1041–1055, 2018.
- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "Multi-Source DOA Estimation through Pattern Recognition of the Modal Coherence of a Reverberant Soundfield", *IEEE/ACM Transactions on Audio Speech and Language Processing*, Volume 28, pp. 605–618, 2019.

Conference Proceedings

- **A. Fahim**, P. N. Samarasinghe, T. D. Abhayapala, and H. Chen, "A planar microphone array for spatial coherence-based source separation", *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, Vancouver, Canada, August 2018.

- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "PSD Estimation of Multiple Sound Sources in a Reverberant Room Using a Spherical Microphone Array", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 76–80, New York, USA, October 2017.
- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "Sound field separation in a mixed acoustic environment using a sparse higher order spherical microphone array", *Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pp. 151–155, San Francisco, USA, October 2017.
- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "Extraction of exterior field from a mixed sound field for 2D height-invariant sound propagation", *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, Xian, China, September 2016.

The following paper also include results from my Ph.D. study, but not included in this thesis:

- P. N. Samarasinghe, H. Chen, **A. Fahim**, and T. D. Abhayapala, "Performance Analysis of a Planar Microphone Array for Three Dimensional Soundfield Analysis", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 249–253, New York, USA, October 2017.

The research work presented in this thesis has been performed jointly with Prof. Thushara D. Abhayapala and Dr. Prasanga N. Samarasinghe. Approximately 80% of this work is my own.

Abdullah Fahim
Research School of Electrical, Energy and Materials Engineering
The Australian National University
Canberra ACT 2601
September 2019

Acknowledgements

I would like to acknowledge the support and guidance of the following people who played pivotal roles in the pursuit of my goal and shaped me to who I am today:

First and foremost, I would like to express earnest gratitude to my Ph.D. supervisor, Professor Thushara D. Abhayapala, who greatly helped and encouraged me at each stage of my Ph.D. candidacy. Without his invaluable guidance, scholarly inputs, and unconditional support, this dissertation would not have been possible. It was absolutely an honour to pursue my Ph.D. under his supervision and to learn from his expertise and vast experience.

I am also sincerely grateful to my co-supervisor, Dr. Prasanga N. Samarasinghe, who was instrumental in my research. Her extensive supervision, continuous motivation and support, insightful suggestions, and affable friendship were extremely helpful and source of my strength during the Ph.D. research.

The Australian National University for the Ph.D. scholarship and funding.

Dr. Juha Merimaa for his guidance and support during my internship at Apple Inc. in California, USA.

My friends and colleagues in the Audio & Acoustic Signal Processing group, especially Xiang, Hanchi, and Aimee for their friendship.

My parents for their eternal love and support, for all the sacrifices they made and patience they showed. My brothers and sisters, Mahfuz, Farzana, Farhana, and Rahik, for their unconditional love and affection. My friends, in-laws and extended family, especially my father-in-law, Mr. Kamal, who assisted and guided me through various obstacles in life.

Finally, I can not thank enough to my wife, Samiya, for her true love and emotional support, being a constant source of joy and happiness, and tolerate my grumpiness in the rough days. I truly appreciate all the sacrifices she made, all the burdens she shared, and all the hurdles she happily faced with me. Thank you for everything!

Abstract

A real-world soundfield is often contributed by multiple desired and undesired sound sources. The performance of many acoustic systems such as automatic speech recognition, audio surveillance, and teleconference relies on its ability to extract the desired sound components in such a mixed environment. The existing solutions to the above problem are constrained by various fundamental limitations and require to enforce different priors depending on the acoustic condition such as reverberation and spatial distribution of sound sources. With the growing emphasis and integration of audio applications in diverse technologies such as smart home and virtual reality appliances, it is imperative to advance the source separation technology in order to overcome the limitations of the traditional approaches.

To that end, we exploit the harmonic decomposition model to dissect a mixed soundfield into its underlying desired and undesired components based on source and signal characteristics. By analysing the spatial projection of a soundfield, we achieve multiple outcomes such as (i) soundfield separation with respect to distinct source regions, (ii) source separation in a mixed soundfield using modal coherence model, and (iii) direction of arrival (DOA) estimation of multiple overlapping sound sources through pattern recognition of the modal coherence of a soundfield.

We first employ an array of higher order microphones for soundfield separation in order to reduce hardware requirement and implementation complexity. Subsequently, we develop novel mathematical models for modal coherence of noisy and reverberant soundfields that facilitate convenient ways for estimating DOA and power spectral densities leading to robust source separation algorithms. The modal domain approach to the soundfield/source separation allows us to circumvent several practical limitations of the existing techniques and enhance the performance and robustness of the system. The proposed methods are presented with several practical applications and performance evaluations using simulated and real-life dataset.

List of Abbreviations

AH	Acoustical holography
ANC	Active noise cancellation
BF	Beamformer
BSS	Blind source separation
CD	Cepstral distance
CNN	Convolutional neural network
CRNN	Convolutional recurrent neural network
DFT	Discrete Fourier transform
DNN	Deep neural network
DOA	Direction of arrival
DRR	Direct to reverberation energy ratio
DSB	Delay and sum beamformer
ESM	Equivalent source method
ESPRIT	Estimation of signal parameters via rotational invariance technique
FFT	Fast Fourier transform
FWSegSNR	Frequency-weighted segmental signal to noise ratio
GCC	Generalised cross-correlation
GSS	Geometric spectral subtraction

HOM	Higher order microphone
ICA	Independent component analysis
LLR	Log-likelihood ratio
LSA	Log spectral estimator
MD	Maximum directivity
MSE	Mean-squared error
MUSIC	Multiple signal classification
MVDR	Minimum variance distortionless response
NAH	Near-field acoustical holography
NMF	Non-negative matrix factorization
NN	Neural network
PESQ	Perceptual evaluation of speech quality
PHAT	Phase transform
PNC	Passive noise cancellation
PSD	Power spectral density
ReLU	Rectified linear unit
RIR	Room impulse response
RMC	Relative modal coherence
RTF	Room transfer function
SIR	Signal to interference ratio
SNR	Signal to noise ratio
SONAH	Statistically optimal near-field acoustical holography
SRMR	Speech to reverberation modulation energy ratio
SRP	Steered response power

SS	Spectral subtraction
STFT	Short-time Fourier transform
STSA	Short-time spectral amplitude estima- tor
SVD	Singular value decomposition
SVM	Support vector machine
TF	Time-frequency
WDO	W-disjoint orthogonality
WF	Wiener filter

Notations and Symbols

The following mathematical notations and symbols are consistent throughout the thesis:

$\mathbb{E}\{\cdot\}$	Expected value
$\lvert\cdot\rvert$	Absolute value
$\lceil\cdot\rceil$	Ceiling opearation
$\lvert\lvert\cdot\rvert\rvert$	Euclidean distance
$(\cdot)^{\dagger}$	Pseudo-inverse operation
$(\cdot)^{*}$	Complex conjugate
$\overline{\overline{(\cdot)}}$	Cardinality of the underlying multiset
$(\cdot)^{T}$	Transpose
$\boldsymbol{x} \cdot \boldsymbol{y}$	Dot product between $\boldsymbol{x}$ and $\boldsymbol{y}$
$x * y$	Convolution between x and y
$\sum_{nm}^{(\cdot)}$	$\sum_{n=0}^{(\cdot)} \sum_{m=-n}^{n}$
$\delta(\cdot)$	Dirac delta function
$\delta_{uv}(\cdot)$	Kronecker delta function
$j_n(\cdot)$	Spherical Bessel function of order n
$J_n(\cdot)$	Bessel function of order n
$h_n(\cdot)$	Spherical Hankel function of order n

$H_n(\cdot)$	Hankel function of order n
$Y_{nm}(\cdot)$	Spherical harmonics of order n and degree m
β_{nm}	Spherical harmonic coefficients of an exterior soundfield
α_{nm}	Spherical harmonic coefficients of an interior soundfield
$\mathcal{R}\{\cdot\}$	Real part of a complex number
$\mathcal{I}\{\cdot\}$	Imaginary part of a complex number
min	Minimum
max	Maximum
k	Wavenumber
c	Speed of sound propagation
∇^2	Laplacian operator
$\{\{\cdot\}\}$	Multiset
$\mathcal{M}_{\mathcal{X}}(\vartheta)$	Multiplicity of ϑ in the multiset $\mathcal{X}$
$\langle\langle\cdot,\ \cdot\rangle\rangle_M$	Concatenate 2 tensors on M^{th} dimension
i	$\sqrt{-1}$
$\int_{\widehat{\boldsymbol{x}}}(\cdot)\ d\widehat{\boldsymbol{x}}$	$\int_0^{2\pi}\int_0^{\pi}(\cdot)\sin(\theta)\ d\theta\ d\phi$

Contents

List of Figures

List of Tables

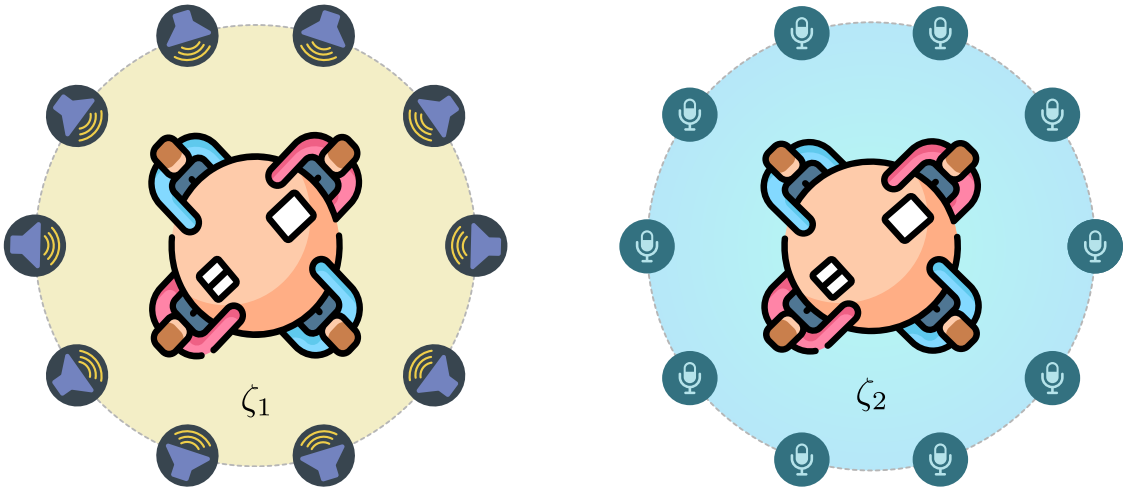
This page intentionally left blank.

Chapter 1

Introduction

1.1 Motivation and Scope

In a real-world scenario, a soundfield is, almost universally, caused by multiple desired and undesired sound sources. In various practical applications such as teleconference, studio/newsroom recording, active noise cancellation (ANC), communication inside an aircraft cockpit, and soundfield reproduction, it is desired, often necessary, to create an isolated sound zone which allows uninterrupted sound recording in complex acoustic scenarios. Furthermore, there are other acoustic signal processing tasks, e.g., audio surveillance, automatic speech recognition, mixing/demixing of music, auditory scene analysis, telecommunication etc., where individual source separation is a prerequisite to achieve a better performance. Historically, source separation techniques mainly involved beamforming, time/frequency domain filtering, and learning-based approaches like independent component analysis (ICA). In the recent years, the spherical harmonics started gaining considerable attention in the audio industry as an attractive tool in various fields of spatial acoustics such as virtual/augmented reality, digital entertainment, ANC technology. This thesis investigates to find intuitive and convenient ways for spatial behaviour modification of a soundfield to perform soundfield separation based on desired source and signal characteristics. Our work mainly focuses on three aspects of a spatial soundfield as described in the following three sections where the first two topics are related to the behavioural modification of a soundfield whereas the last proposition deals



(a) Spatial ANC (b) Spatial capturing

Figure 1.1: Setup for an (a) active noise cancellation system, and (b) soundfield isolation system.

with soundfield prediction based on spatial characteristics.

1.1.1 Zonal separation of a soundfield

Traditionally, the acoustic isolation of a spatial region is achieved by active or passive noise cancellation techniques. ANC targets to eliminate undesired sound using secondary loudspeakers (Fig. 1.1(a)) which requires an accurate estimation of the interfering soundfield. Conversely, passive noise cancellation (PNC) attempts to attenuate the undesired sound by encapsulating the target region with sound absorbing materials which can be inconvenient, often fails to provide desired performance especially at low frequencies, and even unsuitable in certain open environments such as a televised recording. Comparatively, there has not been a lot of progress achieved in capturing 3D soundfield originated from a specific zone in the presence of multiple interfering sources (Fig. 1.1(b)). Williams [1] presented a conceptual theory of scattering near-field holography based on the fundamentals of near-field acoustical holography [2] to isolate the scattering soundfield from the incident waves. Inspired by this methodology, we investigate the prospect of isolating desired and undesired soundfields with respect to distinct source regions from an acoustical point of view and demonstrate its practical applications. Fur-

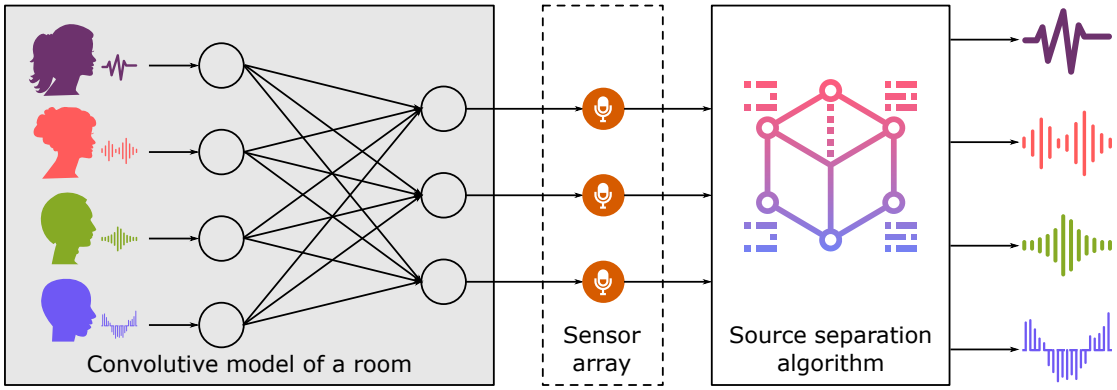


Figure 1.2: A typical flow diagram of a multi-channel source separation task for auditory scene analysis with 3 microphones.

thermore, we envisage the possibility of employing higher order microphones with multi-zone acoustical holography to alleviate logistic complications with respect to array geometry.

1.1.2 Source separation in reverberant environments

For various signal processing tasks, isolating a sound zone may not be enough to achieve the desired outcome, instead a complete separation of individual sound sources becomes necessary. Compared to the zonal separation of a soundfield, source separation (Fig. 1.2) is a relatively well-explored field with various algorithms devised in the literature. A large number solutions have been proposed using learning-based independent component analysis (ICA) [3]–[5] and non-negative matrix factorisation (NMF) [6], [7], however, constrains need to be placed on them to overcome issues like permutation and amplitude ambiguity for ICA and non-convex solution space for NMF. The spatial signal processing utilises a beamformer to boost signal from a certain direction and complement the beamformer output with a suitable post-filtering technique [8]–[11] which stipulates accurate knowledge of power spectral densities (PSD) of the audio components. Hence, source separation still calls for improvement, especially in reverberant environments, and remains an active problem for the researchers. Being a topic closely related to the spatial characteristics of a soundfield, we pursue the solution in the spatial domain by analysing the modal coherence of its spherical harmonic coefficients.

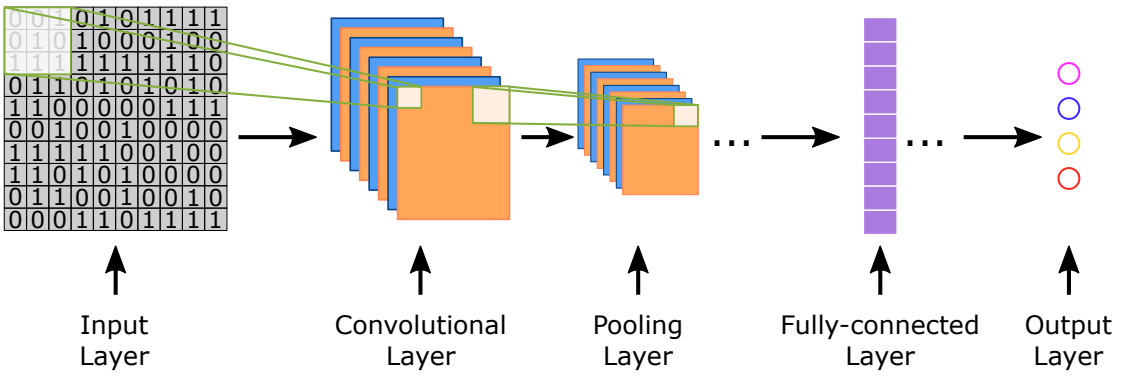


Figure 1.3: Convolutional neural network for image classification.

1.1.3 DOA estimation with multiple overlapping sources

Direction of arrival (DOA) estimation is a well-researched topic in various signal processing tasks such as beamforming, PSD estimation, and spatial coding, which in turn are integral parts of many practical applications. The conventional techniques for DOA estimation such as multiple signal classification (MUSIC) [12] or estimation of signal parameters via rotational invariance technique (ESPRIT) [13] are known to be susceptible to strong reverberation and background noise [14]. The beamformer-based approaches [15] experience degradation in their performance for closely-spaced sources due to the limitation of the spatial resolution. The recent advancements in processing power and storage capacity paves the way for applying deep learning to overcome the limitations of parametric domain implementations of DOA estimation. However, the search for a suitable audio feature to train a neural network is still ongoing and manifest a significant research area. The spherical harmonic decomposition of a soundfield offers a lucrative alternative for such a predictive analysis due to its ability to project a soundfield into space using structured orthogonal basis functions. We use the modal coherence snapshot of a soundfield as a spatial cue to train a convolutional neural network (Fig. 1.3) to learn the unique directional patterns. Furthermore, the same coherence model can be used for both predicting and modifying the spatial behaviour of a soundfield, thus allows resource sharing to achieve an efficient implementation of joint DOA estimation and source separation.

1.2 Problem Statement and Proposed Solution

Based on the discussion in the preceding section, it is evident that the spatial basis functions such as the spherical harmonics open multiple fronts in spatial signal processing to mitigate the limitations of the traditional approaches. Hence, we pose the following research question for us to address in this thesis

How to dissect a soundfield based on source characteristics (e.g., spatial distribution of sources, source directions, and the properties of individual source signals) by projecting the soundfield into space using spatial basis functions?

To this end, we analyse a complex soundfield in the spatial domain by decomposing it utilising spherical harmonics. We extend the concept of acoustical holography into soundfield separation using omni as well as higher order microphones to offer a practical solution of zonal separation of a soundfield. We further investigate the modal coherence of a soundfield to develop mathematical models for multi-source reverberant and noisy environments, which we utilise to estimate PSDs of signal component and demonstrate its application in source separation. The directional dependency of the modal coherence coefficients is exploited to learn the unique patterns as a function of source position to devise an efficient DOA estimation technique using a convolutional neural network. We conduct extensive experimental validations under various practical and simulated environments to measure the performance of the proposed algorithm in various audio processing applications such as:

- **Soundfield separation:** We investigate the zonal soundfield separation up to 1 KHz and achieve an accurate reproduction of the intended soundfield. The method is equally applicable for higher frequencies subject to the availability of the hardware.

- **Accurate PSD estimation:** We solve the modal coherence model from a least-square sense for PSD estimation and demonstrate that the proposed method achieves superior estimation accuracy compared to the competing methods in terms of objective measures as well as visual comparison.

- **Source separation:** The source separation performance is evaluated in different noisy and reverberant real-life environments using a practical setup and measured using industry-standard performance metrics. We also propose a simplified planar array design for source separation and provide an engineering solution to the measurement challenges such as Bessel-zero issue.
- **Efficient DOA estimation:** The modal coherence-based DOA estimation is shown to significantly improve the training efficiency compared to the state-of-the-art techniques. The method offers simplified yet better performing approach for multi-source DOA estimation irrespective of their overlapping nature.

In summary, the proposed solution draws an intuitive representation of soundfield behaviour, explores the existing limitations in spatial audio processing, offers attractive solutions to various essential signal processing problems, and demonstrate better performances compared to the contemporary approaches based on both practical and simulated datasets.

1.3 Thesis Overview and Outline

This thesis offers various tools for spatial analysis and modification of a soundfield to help dissecting it into its primary components. The flow diagram of the thesis outline is shown in Fig. 1.4 where the blocks represent core chapters and are marked with a characteristic symbol $\boldsymbol{P}$, $\boldsymbol{M}$, or $\boldsymbol{E}$ on the top-right corner based on the objective and outcome of the corresponding chapter. A $\boldsymbol{P}$ implies the chapter results in a perceivable outcome whereas $\boldsymbol{M}$s are used with the blocks which are associated in developing underlying mathematical models. The $\boldsymbol{E}$-marked blocks are introduced as an extension to the current workflow to improve system performance. Furthermore, the blocks are inter-connected with a solid or broken arrow. Solid arrows indicate the inter-dependency between the chapters whereas a broken arrow signifies possible route for further improvement that has not been integrated as a part of this thesis.

This thesis produces two perceivable outcomes in the forms of *soundfield separation* and *source separation* (blocks marked with a "P" in Fig. 1.4). The acoustical

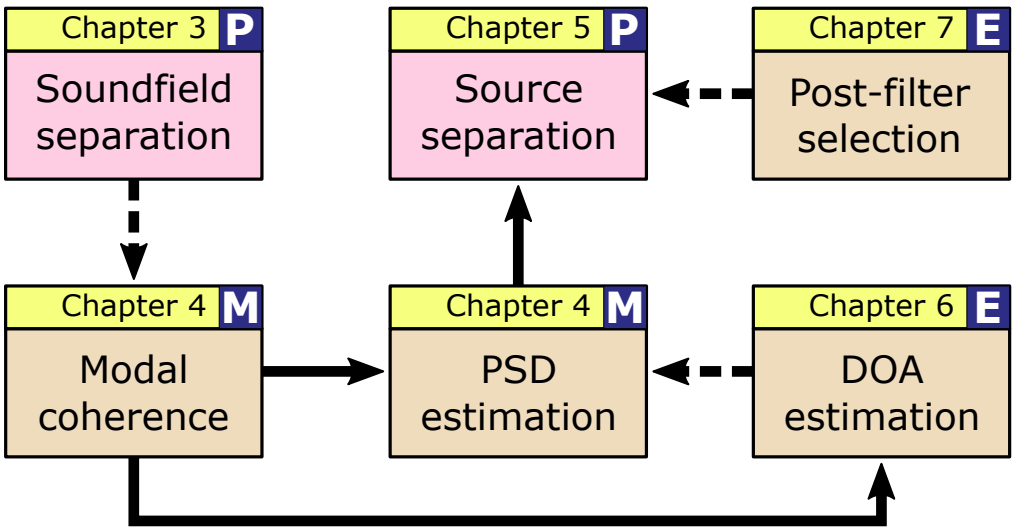


Figure 1.4: Thesis outline. Blocks that produce a perceivable outcome, underlying model, and extension work are marked with $\boldsymbol{P}$, $\boldsymbol{M}$, and $\boldsymbol{E}$, respectively. Solid arrows indicate the inter-dependency between the chapters whereas a broken arrow signifies possible route for further improvement that has not been integrated as a part of this thesis.

holography-inspired *soundfield separation* technique can be used as a standalone application as well as an assist to the modal coherence-based PSD estimation and source separation algorithms (the path is denoted by a broken arrow in Fig. 1.4). Conversely, *source separation* is built on the modal coherence model we develop for noisy and reverberant environments from the spherical harmonic coefficients of a soundfield. A PSD estimator can be devised directly from the modal coherence model and used to support the source separation task. We employ the traditional DOA estimation and post-filtering techniques with our initial proof of concept of source separation, however, later we introduce an efficient DOA estimation technique to overcome some limitations of the traditional approaches. We also make detailed theoretical analysis and comparative study of different post-filtering techniques in order to gain insights into their performance in the presence of non-orthogonal of signals.

The technical contributions of this thesis are distributed into four chapters, as demonstrated in Fig. 1.4. Chapters 3 and 5 produce two perceivable outcomes using the models developed in Chapter 4. Chapter 6 and 7 present two extensions

to complement the core algorithms, as denoted by the broken arrow in Fig. 1.4. The contributions of each chapter of this thesis can be summarised as:

Chapter 2: We provide an extensive literature review of existing works relevant to this thesis and identified the key research gaps we like to address through this work. We also include a brief discussion on the wave propagation model for spherical and cylindrical coordinate systems and review the background theory of spherical harmonic decomposition which is used as the building block for different algorithms we develop in this thesis.

Chapter 3: This chapter devises two separate algorithms based on acoustical holography for separating interior and exterior soundfields from a mixed recording. In the first part of this chapter, we take a dual-surface approach of scattering near-field holography to isolate desired and undesired soundfields for a height-invariant sound propagation model. We then extend the work for 3D sound propagation by developing a multi-zone acoustical holographic model using an array of higher-order microphones (HOM). The latter approach paves the way for a more practical solution from the logistic point of view and increases the robustness of the algorithm due to the inherent properties of HOMs. We also demonstrate practical applications for both the methods in terms of exterior field extraction and speech dereverberation incorporating multiple near-field and far-field sources. The following key outcomes are reached at the end of this chapter:

- The separation technique produces an accurate estimation for soundfield separation and is found to be robust against thermal noise of microphones.
- The proposed method holds the inherent capability of suppressing reverberation and external background noise.
- When the intention is to capture the exterior soundfield, the required number of microphones is less than its theoretical limit.
- The use of HOMs is beneficial for the 3D wave propagation model to reduce the microphone density as well as to increase the robustness against measurement inaccuracies and singularities.

Chapter 4 This chapter analyses the nature of modal coherence in a noisy and reverberant environment. We derive a closed-form mathematical model of modal coherence for a noisy and reverberant soundfield in the presence of multiple concurrent sound sources using their spherical harmonic coefficients. The model coherence devised in this chapter is subsequently used as the framework for PSD estimation. We demonstrate the validity of the model by estimating PSDs in various practical reverberant and noisy environments using real-life dataset utilising a commercially available spherical microphone array. It allows us to measure the robustness of the proposed algorithm against all deviations incurred in a practical environment. In summary, we attain the following major contributions and findings from this chapter:

- Mathematical modelling of noisy and reverberant soundfield using spherical harmonic decomposition.
- Detailed theoretical analysis and demonstration of the practical impact of Bessel-zero issue which, if not addressed in a correct way, significantly limits the performance of a spherical microphone array-based system. We also offer an engineering solution to the Bessel-zero issue.
- The intuition behind the modal power of a reverberant soundfield is explained.
- Comparative performance evaluation and analysis of modal coherence-based PSD estimation is made using a commercially available microphone array under different reverberant and noisy condition.

Chapter 5 In this chapter, we demonstrate a practical application of the modal coherence model by evaluating a source separation performance based on the PSD estimation technique outlined in the previous chapter. Once again, we use the same real-life dataset that incorporates all the practical deviations such as source localisation error, thermal noise at the microphones, non-ideal characteristics of the soundfields etc. We also propose a simpler array structure in the form of a planar array to perform the source separation using a limited hardware support. We

measure the performance using two industry-standard objective metrics, perceptual evaluation of speech quality (PESQ) and frequency-weighted segmental signal to noise ratio (FWSegSNR), and make comparative analysis with contemporary techniques. We show that:

- The proposed algorithm offers significantly better performance compared to the competing methods.
- Based on the performance analysis using oracle and estimated DOAs, we conclude that the algorithm is robust against small DOA estimation error.
- The algorithms are capable of solving under-determined system, hence, it is possible to achieve satisfactory performance using only a subset of the available spherical harmonic coefficients. We exploit this advantage to design a simple planar array for source separation utilising only the even harmonics.
- The condition number of the translation matrix depends on the number of sources as well as the spherical harmonic order.

Chapter 6 Influenced by the optimistic results we obtained from employing the modal coherence model in analysing and modifying soundfield behaviour, in this chapter we explore the idea of predicting spatial characteristics of a soundfield, such as source location, from its modal coherence patterns. We train a convolutional neural network (CNN) to learn the unique attributes of the modal coherence model to determine the source positions. We develop an algorithm to perform multi-source DOA estimation while being trained for only the single-source case irrespective of the overlapping nature of the sources. This allows an efficient and fast training scheme and a seamless run-time performance for DOA estimation in a dynamic scenario where the number of sources vary. We propose a solution to address the occasional violation of W-disjoint orthogonality [16] of the STFT coefficients in a multi-source environment. The algorithm is developed to work independently for azimuth and elevation estimation allowing it to share resources while performing full DOA estimation without affecting its accuracy. We use simulated as well as practical environments to measure the performance of the algorithm and compare

it with a recently proposed CNN-based DOA estimation technique to achieve the following key outcomes:

- The single-source training strategy saves an immense amount of time and resources compared to the contemporary methods.
- The proposed algorithm outperforms the competing method irrespective of the number of sources despite being trained for a single-source scenario.
- Being a data-driven approach, the algorithm is capable of learning the variations in the acoustic setup.
- It achieves more than 98% adjacent accuracy up to 4 sources based on 100 random experiments in different environments, beyond that the performance gradually decreases with approximately 85% adjacent accuracy for 7-source mixture.

Chapter 7 A post-filter at the beamformer output is known to enhance system performance by boosting interference rejection [11]. Hence, it is important to analyse the post-filter design for source separation in the context of this thesis. Several spectral post-filters are available in literature such as Wiener filter, Spectral Subtraction, Log-spectral Amplitude Estimator and so on. These conventional approaches assume orthogonality between the signal components which is occasionally violated due to a limited time-domain support and the short-time stationarity of the speech signals. A recently proposed geometric approach to spectral subtraction (GSS) attempted to resolve this issue, however, GSS imposes additional constraints to compensate for the non-orthogonality. Therefore, a careful consideration for a suitable post-filtering algorithm is required in a complex reverberant environment for achieving a better source separation performance. In this chapter, we review the theory behind the conventional techniques, analyse the constraints and assumptions made, and compare their performance in a practical reverberant and noisy environment. This chapter is intended to act as a working reference for associating a post-filtering technique with various acoustic signal processing tasks such as source separation, dereverberation, and noise suppression. The key findings in this chapter are

- The assumption of orthogonality made in most of the conventional spectral filters is often violated in practice.
- The non-orthogonality is more prominent with reverberation where the reflected signals exhibit a certain level of correlation with the original signal.
- The attempt made in GSS to circumvent the performance issue due to non-orthogonality works only on specific conditions.
- Hence, there exists no global solution, instead the selection should be made based on the application and the nature of the acoustic environment.

Chapter 8 Finally, this chapter provides a summary of the results drawn from this thesis and sheds light on the possible directions for future research.

This page intentionally left blank.

Chapter 2

Literature Review and Background Theory

In this chapter, we review the background theory and current research progresses in the field of soundfield separation based on the source locations. We first explore the literature to survey the state-of-the-art techniques for separating the soundfield contributions by desired and undesired sound sources in a mixed acoustic environment. We then extend our discussion to acoustic source separation in a reverberant environment and analyse the current limitations in the field. An accurate direction of arrival (DOA) estimation of the sound sources is often a prerequisite to both soundfield and source separation. Hence, we draw a comprehensive picture of existing DOA estimation techniques and compare the advantages and limitations of parametric domain with the data-driven approaches. We also identify the major research gaps that we intend to address through this thesis. The last part of this chapter focuses on the background of sound propagation, spherical harmonic decomposition and its properties in the context of this thesis.

2.1 Soundfield Separation over a Large Region

A spatial soundfield is the collective behaviour of sound waves within a region due to one or more sources at various points in space. In many practical applications such as studio/newsroom recording [17], [18], audio surveillance [19], [20], spatial active noise cancellation (ANC) [21], [22], selective soundfield reproduction [23], [24], communication inside an aircraft cockpit [25], [26], and teleconference [27], [28], it is important to estimate the soundfield caused by a subset of the active sound sources. Such a system requires to isolate the desired and undesired soundfields from mixed acoustic measurements in order to capture desired soundfield. A soundfield separation algorithm applies certain spatial audio processing techniques to a mixed soundfield recording in order to isolate the contribution of the desired sound sources from the undesired counterparts. Fig. 2.1 shows a typical example of soundfield recording inside a target region ζ_1 where a blind soundfield separation technique aims to extract the desired soundfield by isolating it from the external interference using microphones placed on the vicinity of the desired zone.

Conversely, soundfield recording has been extensively studied in the literature, especially in the fields of soundfield reproduction and spatial ANC [21], [29]–[34], which is mainly concerned in capturing the soundfield as a whole without making any distinction based on source locations. An interior and exterior soundfield[1] control was proposed in [35] during the reproduction stage to suppress reverberation. The authors of [24] used a combination of source localisation and separation methods and grouped individual sources to achieve selective listening.

2.1.1 Near-field acoustical holography

We apply the principle of acoustical holography in order to achieve blind separation of interior and exterior soundfields without any explicit knowledge about source or room acoustics. Acoustical holography (AH) [36], [37] was originally introduced to reconstruct a 3D soundfield using the measurement on a 2D holographic plane. It uses Rayleigh integrals to solve the forward problem to estimate soundfields away

[1] A soundfield takes the form of an interior or exterior soundfield based on the relative position of the source and the receiver. A detailed definition of interior and exterior soundfield is discussed in Section 2.5.3.

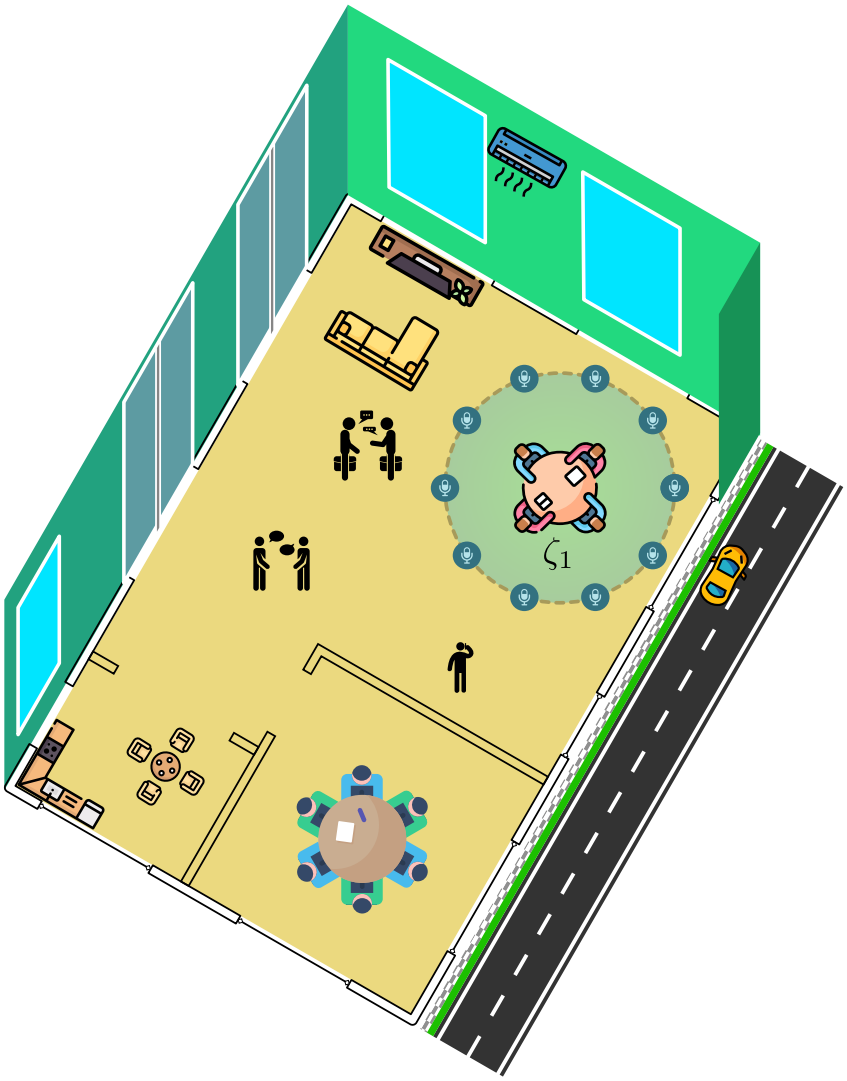


Figure 2.1: Soundfield separation in a mixed acoustic environment.

from the measurement surface and the source region. The spectral resolution of AH is limited by the wavelength of acoustic radiation. To achieve a better spectral resolution and solve the inverse problem of reconstructing soundfield between the measurement plane and the source region, a near-field acoustical holography (NAH) was proposed [2], [38]. Soundfield reconstruction using AH and NAH is based on the assumption that all the sources stay on one side of the measurement plane, i.e. the other side remains source-free. Hence, the traditional NAH fails to distinguish sounds coming from either side of a plane. A solution to this ambigu-

ity was proposed by introducing techniques involving dual measurement surfaces. Measuring the sound pressure or particle velocity, or both, on two closely-spaced planes, it is possible to employ NAH to isolate soundfields caused by sources lying on the opposite side of the measurement surfaces. Since then, the dual surface approach has been adopted in solving various acoustic problems such as measuring reflection coefficients of a test material by observing sound pressure on two parallel planes near the reflective surface [39] or of the ocean bottom [40].

2.1.2 Planar separation of a soundfield

Most of the NAH-based soundfield separation techniques focus on separating sound sources lying on the opposite sides of a 2D plane. Fernandez *et al.* used two closely-spaced rectangular arrays of velocity transducers for separating sources from either side of the arrays [41] based on the contention that NAH with particle velocity produces a better outcome than the conventional sound pressure-based techniques [42]. A further analysis was made in [43] where the authors presented a comparative study between [41] and a joint measurement of sound pressure and particle velocity on a sole measurement surface. One of the major challenges for NAH-based solutions is the mathematical instability due to ill-posed inverse problem in the presence of decaying evanescent waves for which various regularisation strategies have been proposed [44], [45]. Chardon *et al.* argued that the common types of regularisation, e.g., Tikhonov method, often comes with the cost of low spatial resolution for which they tried to overcome using sparse regularisation and compressive sampling techniques [46]. Steiner and Hald proposed a variation of NAH, namely statistically optimal NAH (SONAH) [47], which avoids spatial Fourier transform in order to reduce truncation error caused by a finite-length aperture at the cost of increased computational complexity. Combining the findings from [47] and [42], Jacobsen *et al.* analysed the performance of SONAH based on pressure-velocity probes [48]. Other variations of planar separation of a soundfield include association of NAH with sound pressure-based equivalent source method (ESM) [49], [50], particle velocity-based ESM [51], [52], and wave superposition method [53]. While each of these methods exhibits its strengths and limitations in certain acoustic environments, it is clear that they are not designed to isolate a finite zone surrounded

by undesired sound sources as shown in Fig. 2.1.

2.1.3 Zonal separation of a soundfield

To achieve zonal soundfield separation as depicted in Fig. 2.1, a more appropriate approach is to exploit NAH with cylindrical or spherical coordinate system. Various literary works are available which define NAH or its variant for cylindrical [54]–[57] and spherical [58], [59] geometry. However, it is evident from the foregoing discussion that a sole cylindrical or spherical surface is not enough to separate soundfields originated from either side of the measurement surface. Based on the theory of NAH with the spherical coordinate system, Williams proposed a conceptual design of scattering near-field holography using two concentric circular arrays to isolate the scattering soundfield from the incident waves [1]. We are going to explore this strategy and its practical application in separating interior and exterior soundfields for a height-invariant sound propagation model in the presence of active sound sources on both sides of the surface. It has been established that, for each of concentric circular arrays, at least $(2kr + 1)$ microphones are required [60] to avoid spatial aliasing, where k and r denote the *wavenumber* and radius of the circular array, respectively. Theoretically, it is possible to expand the scattering near-field holography for a 3D soundfield using two concentric spherical arrays with a minimum of $(kr + 1)^2$ microphones in each array. However, this strategy proves to be unrealistic for many practical applications due to a rapid increment of microphone requirement at higher frequencies or with a large target region.

In an acoustic environment where the interior and exterior soundfields coexist, the measured sound pressure in a receiver comprises of a linear mixture of both types of soundfields following the law of superposition. In terms of the spherical harmonic coefficients of a 3D soundfield, the interior coefficients α_{nm} and the exterior coefficients β_{nm} simultaneously contribute to the observed sound pressure. Such a mixture can be decomposed using spherical near-field acoustical holography technique utilising two concentric spherical microphone arrays. However, for a large region, it poses a massive logistic challenge amounting an impractical number of microphones. Hence, it is important to devise an alternative mechanism to the dual surface approach for separating a 3D soundfield. For a more convenient

way of capturing 3D soundfield coefficients over a large region, Samarasinghe *et al.* employed an array of higher order microphones (HOM) [61], [62] which utilises the addition theorem for the Bessel and Hankel functions to relate the local and the global soundfield coefficients. Similar to the original proposition of planar NAH, [62] was also built on the hypothesis of constraining the sources on one side of the spherical surface, thus offering the reconstruction for either interior or exterior soundfield. In this thesis, we apply NAH with HOM-based soundfield decomposition to achieve zonal separation of a soundfield by measuring sound pressure with a sparse HOM array. Furthermore, the application of HOM in estimating spherical harmonic coefficients allows us to use rigid HOMs that improves the robustness of the estimation [63].

A large scale deployment of HOMs in solving practical acoustic problems is still at its early stage despite gaining increasing attention [64]. One of the major reasons for this slow adoption rate is the limited availability of commercially viable hardware. However, in the recent past, a lot of effort has been put into developing new design methodologies, addressing existing limitations, and introducing consumer-friendly and industry-oriented solutions. There have already been a number of commercial HOMs released from different well-known organisations such as "AMBEO VR Mic" from *Sennheiser*, "Eigenmike" from *MH Acoustics*, and "NT-SF1" from *Rodes Microphone*. Furthermore, extensive research is ongoing in the relevant field to improve the quality and scalability of HOMs. Recently, Chen *et al.* proposed a HOM using multiple circular arrays [65], a prototype of which was demonstrated and evaluated in [66]. There also exist several other conceptual designs for HOMs, a few of them are available in the references [67]–[70]. Interested readers are encouraged to explore the theoretical background [71], underlying techniques [63], and fundamental limitations of HOM design such as spatial sampling techniques [72], [73] and soundfield truncation theory [60], [74].

2.2 Acoustic Source Separation Techniques

The last section reviewed the literature associated with the spatial separation of a soundfield. Next, we are going to explore the current progress in the field of source separation and discuss existing limitations and challenges. While a soundfield sep-

aration technique attempts to extract the overall contribution of a group of sound sources enclosed in a region, source separation tracks individual source behaviour in an observed acoustic mixture to recover the underlying source signals. The majority of the source separation methods are proposed using multi-channel audio processing algorithms whereas the single-channel methods predominantly utilise learning-based techniques. In the following sections, we briefly describe a few popular and widely-used source separation algorithms and discuss their strengths and weaknesses.

2.2.1 Blind source separation

A blind source separation (BSS) technique estimates the source signals from mixed observations without any prior knowledge by imposing additional constraints on the source characteristics. Independent component analysis (ICA) is one of the most popular blind source separation techniques that assumes the sources to be statistically independent and having a non-Gaussian distribution. Considering the following blind source mixture model

$$\boldsymbol{x}(n) = \boldsymbol{A}\ \boldsymbol{s}(n) \tag{2.1}$$

given

$$\boldsymbol{x}(n) = \big[x_1(n), x_2(n), \ldots, x_Q(n)\big]$$

$$\boldsymbol{s}(n) = \big[s_1(n), s_2(n), \ldots, s_L(n)\big],$$

where the unknown transfer function $\boldsymbol{A}$ acts on L sound sources $\boldsymbol{s}(n)$ to produce Q observations $\boldsymbol{x}(n)$, ICA jointly estimates $\boldsymbol{A}$ and $\boldsymbol{s}$ exploiting the aforementioned assumptions by defining a cost function such as *Kullback-Leibler* divergence [75] to ensure the maximum statistical independence among the sources, or *negentropy* [76] or *kurtosis* [77] to solve for the most non-Gaussian elements in $\boldsymbol{s}$. ICA requires at least as many microphones as the number of independent sources in the mixture, however, several adaptations of ICA have been proposed to relax this restriction such as sparse signal representation [78] and independent subspace analysis [79]. Further modifications of ICA-based audio source separation were proposed in [80]–[82] to overcome the performance degradation in a reverberant environment due

to convolutive mixing. However, the time-domain approaches to these solutions are computationally expensive while the frequency-domain implementations suffer from well-known *permutation* and *scaling* issues. Hence, despite having several efforts to improve the performance in a reverberant environment, issues remain to be solved in the ICA-based source separation domain.

Non-negative matrix factorisation (NMF) [83] is another genre of BSS that involves in unsupervised decomposition of an audio mixture to extract the underlying sound sources. Unlike ICA, NMF can be formulated for both single and multi-channel cases. The basic idea behind NMF is similar to other matrix factorisation techniques such as single value decomposition (SVD) with the exception that NMF works with, as the name suggests, a non-negative matrix only. Considering $\boldsymbol{X} \in \mathbb{R}^{F \times T}$ being the real-valued spectrum in short-term Fourier transform (STFT) domain, where F and T are total number of frequency and time frames, respectively, NMF acts on $\boldsymbol{X}$ to decompose it into two matrices by minimising a cost function, such as *Kullback-Leibler* divergence or *Frobenius norm*, to obtain

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H} \tag{2.2}$$

where $\boldsymbol{W} \in \mathbb{R}^{F \times K}$ and $\boldsymbol{H} \in \mathbb{R}^{K \times T}$ contain the new basis functions and corresponding time-dependent coefficients, respectively, and K is the reduced dimensionality often coincides with the number of sources. However, NMF does not offer a unique decomposition, rather there exists multiple solutions that satisfy (2.2). Furthermore, without any statistical assumption, clustering of NMF basis functions to represent the underlying source signals appears to be challenging and unreliable. Hence, additional priors are required to be imposed on the solution such as temporal continuity or sparseness of the spectrum [6], [84], [85]. Other variations of NMF to overcome this fundamental limitation include Bayesian extension of NMF which guarantees a convergence of the algorithm by imposing statistical priors [86], a shifted NMF-based clustering algorithm that links each basis with a corresponding note played by an identical instrument [87], and multi-channel extension of NMF for an improved clustering strategy [88].

Comparatively, the history of using deep neural network (DNN) in source separation is relatively short with [89], [90] being among the earliest adoption in this

domain. A DNN is trained for producing a training target for each individual source based on suitable feature snapshots. There are several options available to use as a training target such as ideal binary mask, ideal ratio/soft mask and signal approximation [91]. During the testing stage, the estimated target mask is multiplied with the mixed observation to extract the desired source. A binary mask requires that the feature domain does not contain any overlapping source to avoid producing artefact noise. Conversely, the soft masks work in a similar fashion as other interference rejection filters, however, the algorithm gets complicated and resource intensive with increasing number of sources. Furthermore, the soft masks are generally defined for each individual time-frequency bin, which can be challenging to accurately predict with sparse and non-stationary speech signals. Apart from the selection of the target mask, the choice of input feature also influences the performance of a DNN classifier. Some popular alternatives for feature representation include interaural time and level differences [92], STFT coefficients [93], mel-frequency and gammatone-frequency cepstral coefficient [94], relative spectral transform, and perceptual linear prediction [95]. The choice of neural network can be made arbitrarily or in an empirical manner that suits the target acoustic scenario, a few of which are demonstrated in [96]–[98].

2.2.2 Spatial signal processing-based techniques

The source separation based on spatial filtering majorly deals with array signal processing techniques [99]. The most common approach to this end is beamforming [10], [100] which boosts signal from a particular direction. The selection and design of fixed and adaptive beamformers are well-studied in literature [101], [102], and hence, we skip the discussion in this thesis. Beamforming requires prior knowledge of direction of arrival (DOA) of the desired sources, which is usually acquired along the way using a suitable DOA estimation technique[2]. The performance of a beamforming-based source separation is limited by the directivity index of the beamformer as well as the direction of the interfering sources. In the presence of diffused noise field or reverberation, or narrow spatial separation between sources,

[2]A detailed discussion on the state-of-the-art DOA estimation techniques are presented in the following section

beamforming fails to produce the desired outcome. Hence, a post-filter is commonly associated with a beamformer to complement its output which is known to enhance system performance by boosting interference rejection [11]. Most of the popular post-filtering techniques are modelled in the spectral domain utilising the power spectral densities (PSD) of the individual signal components [103]–[106]. Hence, while the spatial filtering approach attracts a lot of attention due to its simplicity and efficient implementation strategy, it requires a robust and accurate system for PSD estimation in a noisy reverberant conditions.

PSD carries characteristic information of a signal which is useful in various acoustic processing tasks [8], [107], [108]. Several PSD estimation techniques have been proposed in the literature for a single-source in a reverberant environment. Lebart *et al.* [109] used a statistical model of room impulse responses (RIR) to estimate the reverberation PSD and used that in a spectral subtraction-based speech dereverberation technique. Braun *et al.* [110] proposed a PSD estimator using a reference signal under the strict statistical assumption of a diffused reverberant field. Kuklasiński *et al.* [111] developed a maximum likelihood-based method for estimating speech and late reverberation PSD in a single-source noisy reverberant environment assuming a prior knowledge of noise PSD. The spatial correlation between the received microphone signals were utilised in [107] to compute direct to reverberant energy ratio in a noiseless environment by estimating PSDs of the direct and reverberant components. Saruwatari *et al.* [112] devised a method to suppress undesired signals in a multi-source environment using complementary beamformers. A similar idea was adopted in [8] for PSD estimation and source separation utilising multiple fixed beamformers to estimate source PSDs. While [8] is capable of extracting PSDs with a larger number of concurrent sources compared to [112], both the algorithms were developed for the non-reverberant case and presumed known speech and noise source directions. The beamformers used in [8] were chosen empirically which made it vulnerable to ill-posed solutions, hence, a structured design strategy for [8] was explored in [113] exploiting the property of an *M-matrix*.

In the recent years, spherical harmonic decomposition is becoming a popular choice for performing various acoustic processing such as beamforming [114], [115], PSD estimation [116], speech dereverberation [117], and noise suppression [118].

One of the major advantages of spherical harmonic domain representation of a signal is the inherent orthogonality of its basis functions. In our work, we focus towards developing a mathematical model for modal coherence of the spherical harmonic coefficients of a noisy and reverberant soundfield separately and apply that in extracting PSDs of the individual signal components in a multi-source environment.

2.3 Multi-source DOA Estimation

One of the prerequisites of beamforming and PSD estimation is to acquire the knowledge of source locations. Estimating source location can either be restricted to determine the direction of arrival (DOA) only, or finding out the exact source position in space through a source localisation technique. For the purpose of this thesis, it is sufficient to know the DOA of the desired sources, hence, this discussion primarily focuses on existing DOA estimation techniques.

DOA estimation is a decades' old problem with a number of algorithms developed over the years to accurately estimate sound source locations. However, while different algorithms have shown their usefulness under certain environments, they all have their own constraints and limitations and hence, DOA estimation remains an active problem in acoustic signal processing. A large number of DOA estimation techniques have been developed in the parametric domain [119]. There are subspace-based methods like multiple signal classification (MUSIC) [12] or the estimation of signal parameters via rotational invariance technique (ESPRIT) [13] which utilises the orthogonality between the signal and noise subspaces to estimate the source DOAs. MUSIC algorithm was originally developed for narrowband signals, however, it has been extensively used with wideband processing using a frequency smoothing technique [120] or by decomposing the signal into multiple narrowband subspaces [121]. It is common knowledge that the performance of the subspace-based methods are susceptible to strong reverberation and background noise [14]. Recently a variation of MUSIC was proposed in [122] to improve its robustness in a reverberant room assuming the prior knowledge of room coupling coefficients.

There also exist beamforming-based methods for DOA estimation where the

output power of a beamformer is scanned in all possible directions to find out when it reaches the maximum. A popular formulation of the beamformer-based technique is the steered response power (SRP) method which formulates the output power as a sum of cross-correlations between the received signals. Dibiase proposed an improvement to SRP in [15] using the phase transform (PHAT) variant of the generalised cross-correlation (GCC) model [123]. The beamforming-based methods experience degradation in their performance for closely-spaced sources due to the limitation of the spatial resolution. Furthermore, both subspace and beamforming based techniques require to scan for all possible DOA angles during the run time which can be both time and resource intensive. Several modifications have been proposed to reduce the computational cost of SRP-PHAT by replacing the traditional grid search with region-based search [124]–[127], however, this increases the probability of missing a desired source in reverberant conditions.

Another group of parametric approaches to DOA estimation uses the maximum likelihood (ML) optimisation with the statistics of the observed data which usually requires accurate statistical modelling of the noise field [128]–[130]. In more recent works, DOA estimation, posed as a ML problem, was separately solved for reverberant environments [131], [132] and with unknown noise power [133] using expectation-maximisation technique. A large number of localisation techniques are based on the assumption of non-overlapping source mixture in the short-time Fourier transform (STFT) domain, known as W-disjoint orthogonality (WDO) [16]. Li *et al.* adopted a Gaussian mixture model to estimate source locations using ML method on the basis of WDO [134]. The sparsity of speech signals in the STFT domain was exploited in [135], [136] to localise broadside sources by mapping phase difference histogram of the STFT coefficients. The works in [137], [138] imply sparsity on both signals and reflections to isolate time-frequency (TF) bins that contain only direct path contributions from a single source and subsequently estimate source DOAs based on the selected TF bins. Recently, there has been an increase in efforts for intensity-based approaches where both sound pressure and particle velocity are measured and used together for DOA estimation [139]–[142].

Lately, the application of spatial basis functions, especially the spherical harmonics, is gaining researchers' attention in solving a wide variety of acoustic problems including DOA estimation. Among the works we have referred so far in

this thesis, [120], [122], [141], [142] were implemented in the spherical harmonic domain. Tervo *et al.* proposed a technique for estimating DOA of the room reflections based on maximum likelihood optimisation using a spherical microphone array [143]. Kumar *et al.* implemented MUSIC and beamforming-based techniques with the spherical harmonic decomposition of a soundfield for near-field DOA estimation in a non-reverberant environment [115]. A free-field model of spherical harmonic decomposition was used in [144] to perform an optimised grid search for acoustic source localisation. The spherical harmonics are the natural basis functions for spatial signal processing and consequently offers convenient ways for recognising the spatial pattern of a soundfield. Furthermore, the spherical harmonic coefficients are independent of the array structure, hence, the same DOA estimation algorithm can be used with different shapes and designs of sensor arrays as long as they meet a few basic criteria of harmonic decomposition [63], [65]–[68], [71].

Over the past decade, the rapid technology advances in storage and processing capabilities led researchers to lean towards machine learning in solving many practical problems including DOA estimation. Being a data-driven approach, neural networks can be trained for different acoustic environments and source distributions. In the area of single source localisation, significant progresses have been made in solving the limitations in the parametric approaches by incorporating machine learning-based algorithms. The authors of [145]–[147] derived features from different variations of the GCC model to train a neural network for single source localisation. Ferguson *et al.* used both cepstrogram and GCC to propose a single source DOA estimation technique for under-water acoustics [148]. Inspired by the MUSIC algorithm, the authors of [149] utilised the eigenvectors of a spatial correlation matrix to train a deep neural network. Conversely, multi-source localisation poses a more challenging problem to solve, especially with overlapping sources. In the recent past, a few algorithms have been proposed for multi-source localisation based on CNN. A CNN-based multi-source DOA estimation technique was proposed in [150] where the authors used the phase spectrum of the microphone array output as the learning feature. The method in [150] was implemented in the short-time Fourier transform (STFT) domain and all the STFT bins for each time frame were stacked together to form the feature snapshot. On the contrary, Adavanne *et*

al. considered both magnitude and phase information of the STFT coefficients and used consecutive time frames to form the feature snapshot to train a convolutional recurrent neural network (CRNN) and performed a joint sound event detection and localisation [151]. Both [150] and [151] require the model to be trained for unique combinations of sound sources from different angles in order to accurately estimate the DOA of simultaneously active multiple sound sources. Hence, we explore the idea of developing an efficient DOA estimation technique that reduces the resource requirements during training and testing phases as well as offers improved performance compared to the traditional and contemporary techniques by exploiting deterministic nature of the modal coherence model of a soundfield.

2.4 Research Gaps

Based on the foregoing discussion, we identify the following gaps in the existing literature which we aim to address through this thesis:

- **Soundfield separation:** Most of the existing soundfield separation techniques are based on multiple planar surfaces which divide an acoustic zone into two infinite regions, i.e., they slice the soundfield into two halves with respect to the measurement planes. Such a technique is useful in measuring reflection coefficients of a surface [39], [40] or to isolate interfering sources or reflections from a certain direction [48], [51], but does not offer any solution to isolate a bounded region from surrounding external interference. Such a soundfield separation strategy can be deemed attractive in various branches of acoustic signal processing such as teleconference, spatial ANC, soundfield reproduction and so on [17], [20], [22], [25], [28]. Although Williams introduced a conceptual idea of scattering near-field holography using two circular arrays for 2D sound propagation, it was meant for isolating the scattered soundfield from the incident waves and never tested against separating interior and exterior soundfields caused by independent sound sources originated both inside and outside an enclosure.

- **Application of an array of HOM in NAH:** The dual surface strategy of NAH becomes complex and inefficient from a logistic point of view when

we desire to isolate a bounded spatial zone for 3D sound propagation. Due to this limitation, no attempts have been made so far to achieve such an outcome despite having a large number of proposals on planar separation. Furthermore, though we find a handful of work on using a higher order microphone with NAH [59], the contributions and limitations of a HOM array to near-field acoustical holography remains vastly unexplored.

- **Source separation:** Comparatively, we have seen many research works focused on audio source separation over the last few decades. However, although there already exist numerous algorithms for extracting individual sources from an acoustic mixture, the research on source separation is far from over due to the dynamic nature of audio signals. All the existing methods such as ICA, NMF, and beamforming exhibit their individual strengths and limitations under certain environments due to the various assumptions and constraints imposed on them. Hence, the researchers in this field continue to seek better solutions to this problem as the technologies such as deep neural network and soundfield decomposition evolve. We intend to scrutinise the source separation problem from the perspective of spatial basis functions by analysing modal interactions and modelling the coherence between spherical harmonic coefficients of a soundfield. A few of the existing works utilise the spatial correlation matrix in the frequency domain based on measurement data [107] whereas we envisage developing a mathematical model of modal coherence which provides an intuitive explanation of reverberant and noisy soundfield behaviour.

- **DOA estimation:** The advent of fast and efficient machine learning techniques encourages the researchers to rethink the traditional strategies of solving various acoustic problems to overcome their limitations. One of such cases involves DOA estimation where the data-driven approaches are introduced to improve the performance of DOA estimation in challenging acoustic conditions [150], [152]. However, being relatively new research area, DOA estimation using DNN offers lots of scope for improvements in terms of performance and resource efficiency. Furthermore, selecting an intelligent feature for neural networks can significantly boost the performance and reduce algo-

rithm complexity. Being the natural representatives of spatial characteristics, the spherical harmonic coefficients can be an attractive choice for training a machine on the directionality of a soundfield which has not been examined thoroughly. In our attempt to DOA estimation using DNN, we intend to develop a machine learning algorithm that learns the directional characteristics of sound sources based on the unique patterns of modal coherence of its spherical harmonic coefficients.

- **Post-filtering techniques:** The use of post-filtering techniques in speech enhancement and source separation is an old proposition. Most of the existing spectral filters assume orthogonality between different audio components. Recently, it has been shown that this orthogonality assumption does not hold true in the short time frame of STFT domain [153]. Furthermore, in a reverberant condition, the reflected waves maintains a certain level of correlation with the original signal. Hence, it is important to analyse the validity of this assumption in noisy and reverberant environments. While this is a known issue in literature, it lacks a comparative study between the existing methods to find out the extent of deviation due to the violation of orthogonality.

2.5 Harmonic Decomposition of a Soundfield

This thesis makes an extensive use of harmonic decomposition of a soundfield. Most of the theories and models we develop here will be in the modal domain using harmonic coefficients of a soundfield. In this section, we briefly discuss the theory of harmonic representation of a soundfield and outline various spatial audio processing techniques associated with harmonic decomposition.

In audio signal processing, harmonic decomposition expresses a soundfield as a weighted sum of well-defined spatial basis functions to capture, analyse, and/or reproduce diverse acoustic scenarios over a spatial region. This branch of acoustics uses the fundamental solutions of the Helmholtz wave-equation as the basis functions whose nature is principally determined by the wave propagation model. Several techniques have been developed over the years which use various shapes and designs of microphone arrays to estimate the appropriate weights, known as

harmonic coefficients, that characterise a soundfield over a region.

The most common forms of spatial basis functions are spherical and cylindrical harmonics which are inherently orthogonal and can describe a region in terms of a common set of coefficients. The harmonic decomposition offers intuitive and convenient ways for recognising the spatial characteristics of a soundfield which makes it suitable for solving various acoustic problems that involve spatial conundrums. The well-defined orthogonal basis functions allow us to achieve a compact and generalised representation of a soundfield over a sphere. The spherical harmonic coefficients are independent of the array structure, hence, the same harmonic-based algorithm can be used with different shapes and designs of sensor arrays as long as they meet a few basic criteria of harmonic decomposition [63], [65]–[68], [71]. Furthermore, the separation of radial and angular dependencies in harmonic decomposition is proven to be useful in developing various closed-form models and achieving better algorithm efficiency as we have demonstrated throughout this thesis. Lastly, the harmonic decomposition provides conducive means for spatial transformation of a soundfield, such as translation and rotation, using the inherent properties of its basis functions.

In the recent past, the application of harmonic decomposition, especially using the spherical harmonic basis functions, is gaining researchers' attention in solving a wide variety of acoustic problems such as binaural rendering [154]–[157], room acoustic modelling [116], [158]–[160], source localisation [115], [120], [141]–[144], beamforming [72], [115], [161]–[164], soundfield reproduction [31], [165]–[167], active noise cancellation [168]–[171], and dereverberation [117], [172]–[174]. Conversely, cylindrical harmonic decomposition is used for acoustic signal processing involving cylindrical symmetry [175]–[178]. Cylindrical harmonics are also useful in representing a height-invariant soundfield with a simple mathematical model [179], [180] which allows efficient and fast evaluation and often offers a seamless transition to 3D sound propagation model [181]. In this section, we summarise the harmonic decomposition theory and provide a brief overview of different techniques to estimate the harmonic coefficients for spherical and cylindrical wave propagation models.

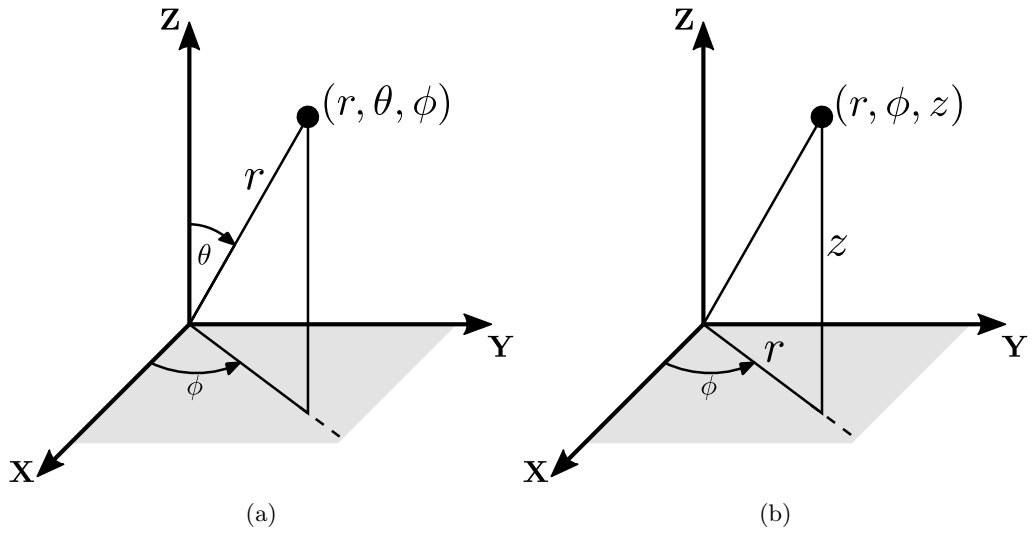


Figure 2.2: The convention used in this work for (a) spherical and (b) Cylindrical coordinate system.

2.5.1 Coordinate system

The mathematical expression for the spatial basis functions varies based on the coordinate system and its convention. Hence, we first define the convention we follow for spherical and cylindrical coordinate system throughout this dissertation.

Fig. 2.2 shows the orientation of the axis and convention we adopt for spherical and cylindrical coordinate systems. The arrows in Fig. 2.2 point towards the positive directions for respective coordinates which follows the right-hand rule. For spherical coordinate system, a point is determined by a triplet (r, θ, ϕ), where radius r is the distance of the point from the origin, elevation θ is the angle formed at the origin by positive Z-axis and the position vector of the corresponding point, and azimuth ϕ is defined as the angle between positive X-axis and the projection of the position vector on XY-plane. On the other hand, the cylindrical coordinate system uses a different triplet (r, ϕ, z) to define a point where the definition of ϕ remains the same with the spherical coordinate system, however, it defines r as the length of the projection of the position vector on XY-plane and utilises the Cartesian parameters z that measures the height of the corresponding point from XY-plane. To define a unique set of coordinates for each point in spherical and

cylindrical coordinate system, we impose the following restrictions in their ranges

$$r \in [0, \infty] \tag{2.3}$$

$$\theta \in [0^\circ, 180^\circ] \tag{2.4}$$

$$\phi \in [0^\circ, 360^\circ]. \tag{2.5}$$

Furthermore, the spherical coordinate (r, θ, ϕ) of a point is related to its Cartesian counterpart (x, y, z) by the following identities

$$(x, y, z) \equiv (r\ \sin\theta\ \cos\phi,\ r\ \sin\theta\ \sin\phi,\ r\ \cos\theta). \tag{2.6}$$

Conversely, the relationship between cylindrical coordinate (r, ϕ, z) and Cartesian coordinate (x, y, z) of the same point is calculated as follows

$$(x, y, z) \equiv (r\ \cos\phi,\ r\ \sin\phi,\ z). \tag{2.7}$$

Letting $\boldsymbol{x}$ be the position vector of a point irrespective of the coordinate system used, we define the sound pressure at $\boldsymbol{x}$ as $p(\boldsymbol{x}, t)$ and $P(\boldsymbol{x}, k)$ in time and frequency domain, respectively, where t is the discrete time index and $k = \frac{2\pi f}{c}$ denotes *wavenumber* with f and c representing frequency and the speed of sound wave, respectively.

2.5.2 Sound propagation in space

In steady state, sound propagates through a homogeneous medium following the wave equation

$$\nabla^2 p(\boldsymbol{x}, t) - \frac{1}{c}\frac{\partial^2 p(\boldsymbol{x}, t)}{\partial t^2} = 0 \tag{2.8}$$

where the spatial Laplacian operator ∇^2 can be expressed based on a chosen coordinate system. The time domain sound pressure $p(\boldsymbol{x}, t)$ is related to the frequency domain representation $P(\boldsymbol{x}, k)$ by the well-known Fourier transform

$$P(\boldsymbol{x}, k) = \int_{-\infty}^{\infty} p(\boldsymbol{x}, t) e^{-i2\pi f t} \tag{2.9}$$

where $i = \sqrt{-1}$. For the purpose of this thesis, we are mainly interested in harmonic decomposition of $P(\boldsymbol{x}, k)$ using complex harmonics. Consequently, we focus on the Fourier transform of (2.8), known as the Helmholz equation, as

$$\nabla^2 P(\boldsymbol{x}, k) + k^2 P(\boldsymbol{x}, k) = 0. \tag{2.10}$$

Equation (2.10) is the basis of all the harmonic analysis discussed in this thesis. In the following sections, we are going to discuss the cylindrical and spherical harmonic decomposition by solving the Helmholz equation for corresponding model of wave propagation.

2.5.3 Harmonic decomposition of a soundfield

We achieve the harmonic decomposition of a soundfield by solving (2.10) for appropriate coordinate system. The solution approach uses a technique called *separation of variables* which assumes that the general solution can be written as a product of independent solutions for each coordinate.

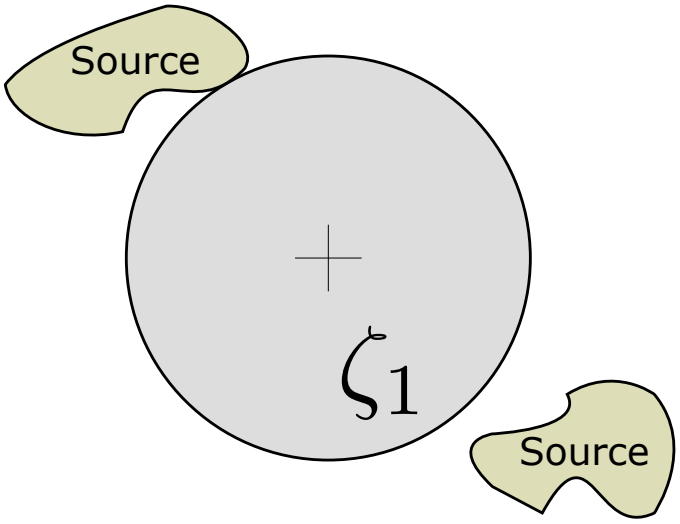
Spherical harmonic decomposition

We use the appropriate definition of ∇^2 for the spherical coordinate system and apply *separation of variables* technique on (2.10) to reach the alternative solutions of the wave equation in a 3D space as [1, ch. 6]

$$P(\boldsymbol{x}, k) = \sum_{nm}^{\infty} \Big[\alpha_{nm}(k)\ j_n(kr) + \alpha_{nm}^{(2)}(k)\ y_n(kr)\Big] Y_{nm}(\widehat{\boldsymbol{x}}) \tag{2.11}$$

$$P(\boldsymbol{x}, k) = \sum_{nm}^{\infty} \Big[\beta_{nm}(k)\ h_n(kr) + \beta_{nm}^{(2)}(k)\ h_n^{(2)}(kr)\Big] Y_{nm}(\widehat{\boldsymbol{x}}) \tag{2.12}$$

where $\boldsymbol{x} \equiv (r, \theta, \phi)$ for spherical coordinate system, $\sum_{nm}^{\infty} \equiv \sum_{n=0}^{\infty} \sum_{m=-n}^{n}$, α and β are known as spherical harmonic coefficients of a soundfield, $j_n(\cdot)$ and $y_n(\cdot)$ are the spherical Bessel functions of the first and second kind, respectively, and $h_n(\cdot)$ and $h_n^{(2)}(\cdot)$ are the spherical Hankel functions of the first and second kind, respectively.



(a)

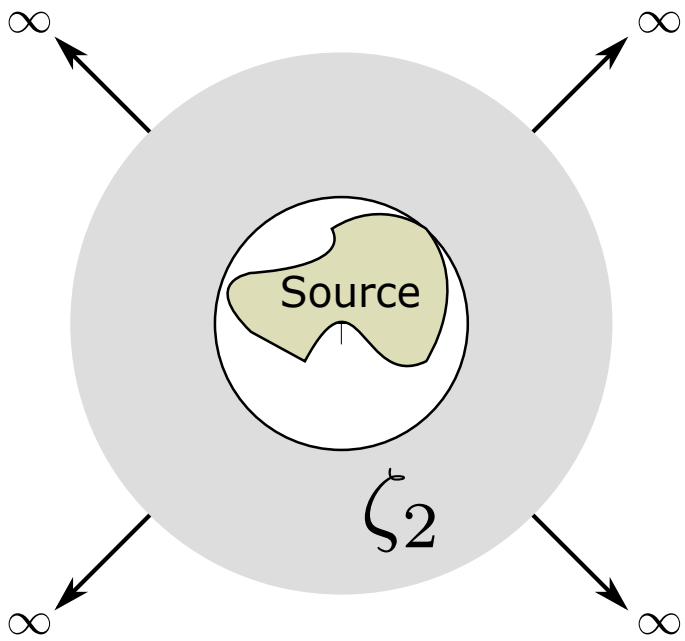
Figure 2.3: Examples of an interior soundfield, ζ_1.

The complex spherical harmonics $Y_{nm}(\widehat{\boldsymbol{x}})$ is defined as

$$Y_{nm}(\widehat{\boldsymbol{x}}) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}}\,\mathcal{P}_{n|m|}\big(\cos\theta\big)\;e^{im\phi} \tag{2.13}$$

where $\widehat{\boldsymbol{x}} \equiv (\theta, \phi)$, $|\cdot|$ denotes absolute value, $(\cdot)!$ represents factorial, and $\mathcal{P}_{n|m|}(\cdot)$ is an associated Legendre polynomial. Both (2.11) and (2.12) offers solutions to the wave equation (2.10) in a source-free region, however, (2.11) is more appropriate for standing waves whereas (2.12) is considered suitable in explaining travelling wave equations.

The categorisation of a soundfield based on the relative positions of sound sources has been extensively studied in literature [1], [71], [165], [182], [183]. For an interior soundfield, the sound sources are located entirely outside the region of validity ζ_1, as shown in Fig. 2.3. For mathematical tractability, we define the centre of ζ_1 as the global origin which implies we must have a finite sound pressure at origin. Based on the fact that the Hankel functions approach infinity at origin, we must exclude (2.12) from the set of possible solutions for an interior soundfield. Furthermore, the spherical Bessel function $y_n(\cdot)$ is also infinite at origin, hence, we set $\alpha^{(2)}_{nm}(\cdot) = 0$ in (2.11) to get the spherical harmonic decomposition of an interior



(b)

Figure 2.4: Examples of an exterior soundfield, ζ_2.

soundfield as

$$P_I(\boldsymbol{x}, k) = \sum_{nm}^{\infty} \alpha_{nm}(k)\ j_n(kr)\ Y_{nm}(\widehat{\boldsymbol{x}}) \tag{2.14}$$

where the subscript $(\cdot)_I$ indicates an interior soundfield.

Conversely, an exterior soundfield is demonstrated in Fig. 2.4 where the region of validity ζ_2 lies beyond the farthest point of the source region with respect to the origin. Since the origin is not included in the region of validity, we can use (2.12) as the solution to an exterior soundfield. Furthermore, the asymptotic behaviour of the Hankel functions dictates that $h_n^{(2)}(kr)$ represents an incoming wave which cannot be true for an exterior field. Hence, the appropriate solution for such a case is reached by letting $\beta_{nm}^{(2)}(\cdot) = 0$ in (2.12), i.e.,

$$P_E(\boldsymbol{x}, k) = \sum_{nm}^{\infty} \beta_{nm}(k)\ h_n(kr)\ Y_{nm}(\widehat{\boldsymbol{x}}) \tag{2.15}$$

where $P_E(\cdot)$ denotes sound pressure due to an exterior soundfield.

Cylindrical harmonic decomposition

Based on the application, the solution of wave equation in the cylindrical coordinate system can prove to be more desirable in certain areas. Cylindrical harmonic decomposition is of great interest in the fields that extensively use cylindrical symmetry such as in near-field holography [1], submarine vibration and radiation patterns analysis [176], and noise control in a circular duct [177]. In this thesis, we use reduced cylindrical harmonic decomposition when we assume height-invariant sound propagation model for mathematical simplicity.

Following a similar technique as in the last section, we can reach the solutions to the wave equation (2.10) using cylindrical coordinates as [1, ch. 4]

$$P(\boldsymbol{x},k)=\sum_{n=-\infty}^{\infty}e^{in\phi}\int_{-\infty}^{\infty}\Big[\alpha_n(k,k_z)\;J_n(k_r r)+\alpha_n^{(2)}(k,k_z)\;Y_n(k_r r)\Big]e^{ik_z z}\;dk_z \quad (2.16)$$

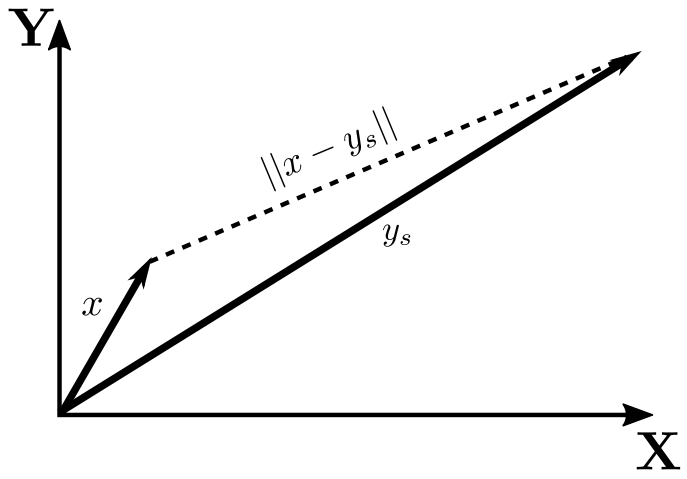
$$P(\boldsymbol{x},k)=\sum_{n=-\infty}^{\infty}e^{in\phi}\int_{-\infty}^{\infty}\Big[\beta_n(k,k_z)\;H_n(k_r r)+\beta_n^{(2)}(k,k_z)\;H_n^{(2)}(k_r r)\Big]e^{ik_z z}\;dk_z \quad (2.17)$$

where $\boldsymbol{x}\equiv(r,\phi,z)$ for cylindrical coordinate system, $k_r=\sqrt{k^2-k_z^2}$, α and β are cylindrical harmonic coefficients of a soundfield, $J_n(\cdot)$ and $Y_n(\cdot)$ are the Bessel functions of the first and second kind, respectively, and $H_n(\cdot)$ and $H_n^{(2)}(\cdot)$ are the Hankel functions of the first and second kind, respectively. Equation (2.16) and (2.17) are appropriate for standing and travelling waves, respectively. In the context of this thesis, we are mostly interested in the height-invariant wave propagation where $\frac{\partial^2 P(\boldsymbol{x},k)}{\partial z^2}=0$ and thus $k_z=0$, hence, (2.16) and (2.17) reduce to

$$P(\boldsymbol{x},k)=\sum_{n=-\infty}^{\infty}\Big[\alpha_n(k)\;J_n(kr)+\alpha_n^{(2)}(k)\;Y_n(kr)\Big]e^{in\phi} \quad (2.18)$$

$$P(\boldsymbol{x},k)=\sum_{n=-\infty}^{\infty}\Big[\beta_n(k)\;H_n(kr)+\beta_n^{(2)}(k)\;H_n^{(2)}(kr)\Big]e^{in\phi}. \quad (2.19)$$

Henceforth, using the boundary conditions for interior and exterior soundfield, we can deduce the cylindrical harmonic decomposition of height-invariant soundfield



(a)

Figure 2.5: Examples of a near-field sound source.

as

$$P_I(\boldsymbol{x}, k) = \sum_{n=-\infty}^{\infty} \alpha_n(k)\ J_n(kr)\ e^{in\phi} \tag{2.20}$$

$$P_E(\boldsymbol{x}, k) = \sum_{n=-\infty}^{\infty} \beta_n(k)\ H_n(kr)\ e^{in\phi} \tag{2.21}$$

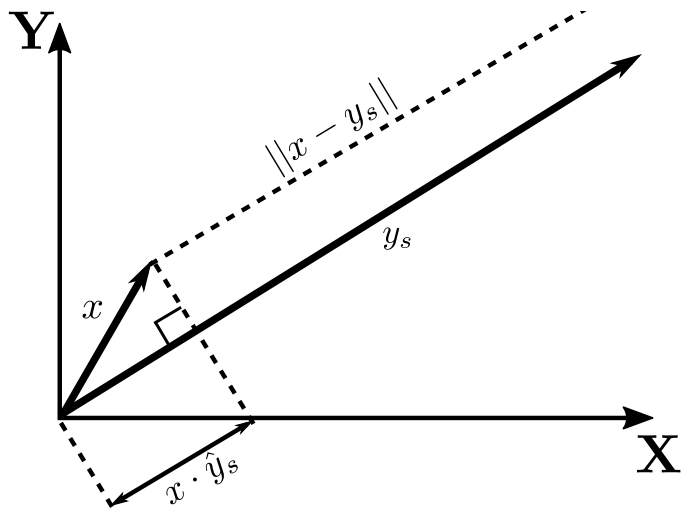
where (2.20) and (2.21) represents interior and exterior soundfields, respectively.

In the next section, we review the analytical expression of harmonic coefficients of a soundfield caused by near-field and far-field sound sources.

Analytical expression of harmonic coefficients of a soundfield

Let us consider $P(\boldsymbol{x}|\boldsymbol{y}_s, k)$ be the sound pressure at $\boldsymbol{x} \equiv (r, \theta, \phi) \equiv (r, \widehat{\boldsymbol{x}})$ caused by a near-field source at $\boldsymbol{y}_s \equiv (r_s, \theta_s, \phi_s) \equiv (r_s, \widehat{\boldsymbol{y}}_s)$ (Fig. 2.5). It can be shown that $P(\boldsymbol{x}|\boldsymbol{y}_s, k)$, called the Green's function, satisfies the homogeneous Helmholz equation (2.10) everywhere except at $\boldsymbol{x} = \boldsymbol{y}_s$ [184, pp. 25-26]. Mathematically, the Green's function is the solution to the inhomogeneous Helmholz equation

$$(\nabla^2 + k^2)P(\boldsymbol{x}|\boldsymbol{y}_s, k) = -\delta(\boldsymbol{x} - \boldsymbol{y}_s) \tag{2.22}$$



(b)

Figure 2.6: Examples of a far-field sound source.

where $\delta(\cdot)$ is a Dirac delta function. It is shown in [184, pp. 26] that the solution to (2.22) is

$$P(\boldsymbol{x}|\boldsymbol{y}_s, k) = \frac{e^{ik\|\boldsymbol{x}-\boldsymbol{y}_s\|}}{4\pi \|\boldsymbol{x}-\boldsymbol{y}_s\|} \tag{2.23}$$

where $\|\cdot\|$ denotes Euclidean norm. This can be expanded into spherical harmonic domain using the addition theorem [185, pp. 31-32] as

$$\frac{e^{ik\|\boldsymbol{x}-\boldsymbol{y}_s\|}}{4\pi \|\boldsymbol{x}-\boldsymbol{y}_s\|} = \begin{cases} \sum_{nm}^{\infty} ik\ h_n(kr_s)\ Y_{nm}^*(\widehat{\boldsymbol{y}}_s)\ j_n(kr)\ Y_{nm}(\widehat{\boldsymbol{x}}), & r < r_s \\ \sum_{nm}^{\infty} ik\ j_n(kr_s)\ Y_{nm}^*(\widehat{\boldsymbol{y}}_s)\ h_n(kr)\ Y_{nm}(\widehat{\boldsymbol{x}}), & r > r_s \end{cases} \tag{2.24}$$

where $(\cdot)^*$ denotes complex conjugate. Clearly, the first and second case of (2.24) satisfy the boundary condition of an interior and exterior soundfield, respectively. Hence, comparing (2.24) with (2.14) and (2.15), we get

$$\alpha_{nm}(k) = ik\ h_n(kr_s)\ Y_{nm}^*(\widehat{\boldsymbol{y}}_s) \tag{2.25}$$

$$\beta_{nm}(k) = ik\ j_n(kr_s)\ Y_{nm}^*(\widehat{\boldsymbol{y}}_s). \tag{2.26}$$

When $r_s \gg r$, the source is considered a far-field source and the sound propagates as a plane wave. Under such assumption, the denominator of the Green's function in (2.23) can be reduced by approximating $\|\boldsymbol{x} - \boldsymbol{y}_s\| \approx r_s$. However, due to the oscillating nature of the frequency-dependent exponential, we can not use the same approximation with the numerator in the right hand side of (2.23). Instead, we use a more accurate approximation $\|\boldsymbol{x} - \boldsymbol{y}_s\| \approx r_s - \boldsymbol{x} \cdot \widehat{\boldsymbol{y}}_s$ [185, pp. 21], which is also evident from Fig. 2.6, to get a far-field approximation of the Green's function

$$P(\boldsymbol{x}|\boldsymbol{y}_s, k) = \frac{e^{ik(r_s - \boldsymbol{x}\cdot\widehat{\boldsymbol{y}}_s)}}{4\pi r_s}. \tag{2.27}$$

Furthermore, the asymptotic behaviour of the spherical Hankel functions for large argument suggests [185, pp. 30]

$$h_n(kr_s) = (-i)^{n+1}\ \frac{e^{ikr_s}}{kr_s}, \quad r_s \to \infty. \tag{2.28}$$

Hence, by using (2.27) and (2.28) in (2.24) and then cancelling the common terms from both sides, we get the spherical harmonic expansion of the far-field approximation of the Green's function

$$e^{-ik\ \boldsymbol{x}\cdot\widehat{\boldsymbol{y}}_s} = \sum_{nm}^{\infty} 4\pi\ (-i)^n\ Y^*_{nm}(\widehat{\boldsymbol{y}}_s)\ j_n(kr)\ Y_{nm}(\widehat{\boldsymbol{x}}), \quad r_s \gg r. \tag{2.29}$$

Consequently, comparing (2.29) with (2.14), we get the spherical harmonic coefficients for a far-field source as

$$\alpha_{nm}(k) = 4\pi\ (-i)^n\ Y^*_{nm}(\widehat{\boldsymbol{y}}_s). \tag{2.30}$$

Note that, for mathematical tractability, (2.24) and (2.29) are normalised with respect to the sound pressure at $\boldsymbol{y}_s$ and origin, respectively. If we have to ensure the same scaling for near-field and far-field sources, we need to keep the common terms on both sides of (2.29) such that the left hand side of (2.29) matches (2.27).

In a similar manner, the cylindrical harmonic expansion of the Green's function for 2D height-invariant sound propagation can be used to show that the cylindrical harmonic coefficients due to a near-field source is given by [186, pp. 13-15] [1, Ch.

4] [187]

$$\alpha_{nm}(k) = H_n(kr_s)\ e^{-in\phi_s} \tag{2.31}$$

$$\beta_{nm}(k) = J_n(kr_s)\ e^{-in\phi_s} \tag{2.32}$$

where a point source is located at (r_s, ϕ_s, z_s) in cylindrical coordinate system. Finally, the cylindrical harmonic coefficients for an interior height-invariant soundfield due to a far-field source is given by [186, pp. 14]

$$\alpha_{nm}(k) = (-i)^n\ e^{-in\phi_s}. \tag{2.33}$$

2.5.4 Properties of spherical harmonics

A large part of this thesis deals with various properties of spherical harmonics, a few of which are described subsequently.

Spherical harmonics are a set of orthonormal basis functions which can represent a continuous function defined over a sphere. Mathematically, we can express the orthonormality of spherical harmonics by

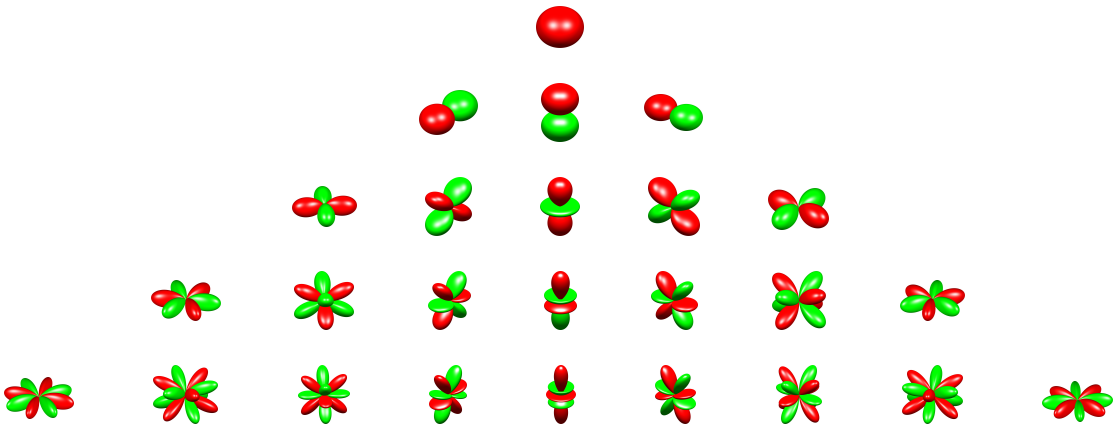
$$\int_{\widehat{\boldsymbol{x}}} Y_{nm}(\widehat{\boldsymbol{x}})\ Y^*_{n'm'}(\widehat{\boldsymbol{x}})\ d\widehat{\boldsymbol{x}} = \delta_{nn'}\delta_{mm'} \tag{2.34}$$

where the Kronecker delta function δ_{ab} is defined as

$$\delta_{ab} = \begin{cases} 1, & a = b \\ 0, & a \neq b. \end{cases} \tag{2.35}$$

Each spherical harmonic mode $Y_{nm}(\cdot)$ is associated with an order n and degree m and favours certain directionality. The orders of the spherical harmonics are non-negative integers and can be extended infinitely. However, in most of the practical cases, a first few orders of the spherical harmonics are enough to represent a soundfield with acceptable accuracy [60], [166]. On the other hand, m is closely related to n and can vary in the range of $[-n, n]$. Fig. 2.7 shows all the spherical harmonics up to 4^{th} order.

Due to the recurrent property of the associated Legendre functions, the spherical



(b)

Figure 2.7: Spherical harmonics up to 4^{th} order.

harmonics exhibit the following symmetry [1, pp. 191]

$$Y^*_{nm}(\widehat{\boldsymbol{x}}) = (-1)^m \; Y_{n(-m)}(\widehat{\boldsymbol{x}}). \tag{2.36}$$

The orthonormal spherical harmonics of order n obeys the addition theorem [185, pp. 27-28] which is often used in simplifying harmonic representation of a soundfield

$$\sum_{m=-n}^{m=n} Y_{nm}(\widehat{\boldsymbol{x}}) \; Y^*_{nm}(\widehat{\boldsymbol{y}}) = \frac{2n+1}{4\pi} \mathcal{P}_n\big(\cos\theta_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{y}}}\big) \tag{2.37}$$

where $\mathcal{P}_n(\cdot)$ is Legendre polynomial and $\theta_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{y}}}$ denotes the angle between $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{y}}$. Another important property that we extensively use in this thesis is the behaviour of the integration of three different spherical harmonics over a sphere [188, pp. 63]

$$\int_{\widehat{\boldsymbol{x}}} Y_{nm}(\widehat{\boldsymbol{x}}) \; Y_{n'm'}(\widehat{\boldsymbol{x}}) \; Y_{n''m''}(\widehat{\boldsymbol{x}}) \; d\widehat{\boldsymbol{x}} = \sqrt{\frac{(2n+1)(2n'+1)(2n''+1)}{4\pi}} \begin{pmatrix} n & n' & n'' \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} n & n' & n'' \\ m & m' & m' \end{pmatrix} \tag{2.38}$$

where $(\cdot)$ in (2.38) represents Wigner-3j symbol [189]. The integral property of the spherical harmonics is very useful in describing a defused soundfield due to the fact that the integration becomes space-independent.

We use the aforementioned properties of the spherical harmonics along with the harmonic decomposition of a soundfield in achieving various desired spatial modification.

2.5.5 Estimating harmonic coefficients from measurements

We have discussed the synthesis equation to reconstruct a soundfield from its harmonic coefficients. We also have shown the analytical expression of the coefficients for a unit impulse in a non-reverberant environment. However, in a practical situation, often we face scenarios where a soundfield is composed of time-varying sources and the reflections from surrounding walls and objects. In such a case, we need to rely on the measured sound pressure or velocity to estimate the harmonic coefficients instead of using the aforementioned analytical solutions. A large number of array design and algorithms are proposed in literature related to extracting soundfield coefficients based on measurements [63], [65], [67]–[71]. Ideally for the same acoustic setup, the harmonic coefficients should remain constant irrespective of the underlying techniques. However, the values can vary slightly depending on the accuracy of the design and extraction procedures.

In this section, we briefly summarise the estimation of soundfield coefficients using a spherical microphone array [63], [71]. Multiplying both sides of (2.14) by $Y_{n'm'}(\widehat{\boldsymbol{x}})$ and integrating on the surface of a sphere, we get

$$\int_{\widehat{\boldsymbol{x}}} P_I(\boldsymbol{x}, k)\; Y_{n'm'}(\widehat{\boldsymbol{x}})\; d\widehat{\boldsymbol{x}} = \sum_{nm}^{\infty} \alpha_{nm}(k)\; j_n(kr) \int_{\widehat{\boldsymbol{x}}} Y_{nm}(\widehat{\boldsymbol{x}})\; Y_{n'm'}(\widehat{\boldsymbol{x}})\; d\widehat{\boldsymbol{x}}. \tag{2.39}$$

Using the orthogonal property of spherical harmonics from (2.34) in (2.39), we obtain

$$\alpha_{nm}(k) = \frac{1}{j_n(kr)} \int_{\widehat{\boldsymbol{x}}} P_I(\boldsymbol{x}, k)\; Y_{n'm'}(\widehat{\boldsymbol{x}})\; d\widehat{\boldsymbol{x}}. \tag{2.40}$$

Eq. (2.40) poses an impractical requirement of a continuous aperture over a sphere to estimate $\alpha_{nm}(k)$. Instead, spatial sampling techniques [73] are utilised to ap-

proximate (2.40) with

$$\alpha_{nm}(k) \approx \frac{1}{j_n(kr)} \sum_{q=1}^{Q} w_q \; P_I(\boldsymbol{x}_q, k) \; Y^*_{nm}(\widehat{\boldsymbol{x}}_q) \tag{2.41}$$

where Q is the number of microphones and $w_q \; \forall q$ are corresponding microphone weights, determined by the spatial sampling scheme, in order to enforce the orthonormal property of the spherical harmonics with a limited number of sampling points, i.e.

$$\sum_{q=1}^{Q} w_q \; Y_{nm}(\widehat{\boldsymbol{x}}_q) \; Y^*_{n'm'}(\widehat{\boldsymbol{x}}_q) \approx \delta_{nn'}\delta_{mm'}. \tag{2.42}$$

Several array structures have been proposed in the literature to estimate α_{nm} using (2.41) such as a spherical open/rigid array [63], [71], multiple circular arrays [68], or a planar array with differential microphones [65]. For a general case, $j_n(kr)$ in (2.41) is replaced by $b_n(kr)$ such that [1, pp. 228-230]

$$b_n(\xi) = \begin{cases} j_n(\xi) & \text{for an open array} \\ j_n(\xi) - \frac{j'_n(\xi)}{h'_n(\xi)} h_n(\xi) & \text{for a rigid spherical array} \end{cases} \tag{2.43}$$

where $\xi \in \mathbb{R}$ and $(\cdot)'$ refers to the corresponding first derivative term.

2.6 Summary

In this chapter, we performed a comprehensive literature survey on three distinct acoustic challenges - (1) spatial separation of a soundfield, (2) source separation in reverberant environment, and (3) source localisation and DOA estimation for single and multi-source environments. We also discussed background theories of sound propagation and signal representation techniques that we utilise in this thesis to devise solutions to the aforementioned problems. We identified the existing gaps in literature that we intend to address in this thesis. We also laid down the foundation of spherical harmonics and its contribution in soundfield decomposition which is used in the subsequent chapters to analyse, predict, and modify spatial characteristics of a soundfield. In the next chapter, we devise multiple algorithms

to isolate interior and exterior soundfields from an acoustic mixture utilising the harmonic decomposition in spherical and cylindrical coordinate systems. We further investigate the characteristics of a reverberant soundfield in Chapter 4 and propose a novel algorithm for PSD estimation establishing a closed-form expression of modal coherence between the spherical harmonic coefficients. Chapter 5 demonstrates a practical application of the PSD estimation technique in terms of source separation using full and reduced coherence matrix. The uniqueness of the modal coherence patterns with respect to source directions are analysed in Chapter 6 to train a convolutional neural network for multi-source DOA estimation with an efficient training and evaluation strategy. Finally in Chapter 7, we analyse limitations of different post-filtering methods under various practical environments in an attempt to understand the acoustic bottleneck of the source separation algorithm.

This page intentionally left blank.

Chapter 3

Soundfield Separation over a Large Spatial Region

A soundfield separation technique decomposes a mixed soundfield into multiple segments based on the spatial location of the inducing sound sources. Soundfield separation over a large region holds significant potentials to solve various acoustic problems in sound recording, reproduction, and noise cancellation fields. This chapter explores the practical viability of soundfield separation over a bounded region and studies its strengths and limitations in different simulated environments. We first consider a height-invariant sound propagation to validate the concept in a docile environment and analyse its outcome. We then expand the methodology for a 3D sound propagation model and devise an alternative theory for an efficient soundfield separation using higher order microphones in near-field acoustical holography. Both the cases are complemented by multiple experimental evaluations based on different acoustic scenarios.

3.1 Introduction

A soundfield caused by one or more sound sources takes the form of an interior, exterior or a mixed soundfield based on the relative locations of the sound sources.

The recording and reproduction of a combined soundfield has been extensively studied in the literature [21], [29], [31], [33], however, the challenging task of separating them from mixed measurements to achieve selective recording remains largely unexplored. But in nature, the interior and exterior soundfields often co-exist, hence, their isolation is key to the success in various branches of acoustic signal processing such as audio surveillance, spatial active noise cancellation, selective soundfield reproduction and so on [17], [20], [21], [25], [28]. The planar separation of soundfield has been discussed and analysed in the past [41], [42], [45], [47] using near-field acoustical holography [2], however, it offers only a limited scope in practical application, e.g., measuring reflection coefficients of a testing material [39]. Comparatively, not much has been explored in the more generic case of separating soundfield from/inside a bounded region. To the best of our knowledge, the conceptual design of scattering near-field holography to isolate the incident and reflected waves [1] is the nearest topic in this area. However, [1] poses significant challenges for its application with 3D soundfield due to the requirement of a large number of microphones which affects the practical feasibility as well as the robustness of the algorithm.

In this Chapter, we extend the concept of scattering near-field holography to isolate two independent acoustic zones, both containing active sound sources, to extract the interior and exterior soundfields based on the measurements on two holographic circular planes for 2D height-invariant sound propagation. We evaluate the performance of this method in separating a desired soundfield from the interfering sources as well as demonstrate its applicability in traditional speech enhancements such as speech dereverberation. We then extend the multi-zone holographic technique to 3D sound propagation utilising an array of higher order microphones to improve the performance, robustness, as well as practical viability. The proposed solution considers a sparse distribution of the higher order microphones which further reduces the design complexity by offering a simpler geometry. Both the techniques are evaluated under different noisy and noiseless environments to measure the effectiveness as well as the robustness of the algorithms.

The rest of the chapter is organised as follows. Section 3.2 introduces the problem statement and develop a soundfield separation technique for height-invariant sound propagation. The problem is redefined for a 3D soundfield in Section 3.3

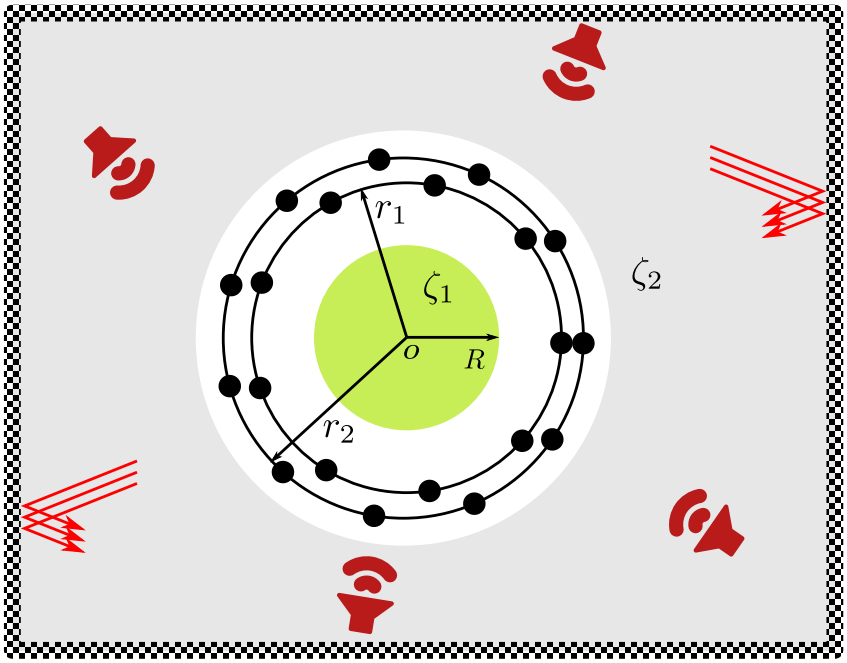


Figure 3.1: Room geometry for a 2D height-invariant sound propagation model.

and a novel algorithm is proposed using an array of higher order microphones envisioning a more robust and viable solution for a practical scenario. Both the models are scrutinised under different acoustic environments and the experimental results are presented in Section 3.4. Finally, a conclusion is drawn in Section 3.5 with an indication of the forthcoming work.

3.2 Height-invariant Sound Propagation

This section considers a special scenario of height-invariant sound propagation model. We simultaneously measure sound pressure on two circular planes and utilise near-field acoustical holography based on a cylindrical coordinate system to accomplish zonal separation of a soundfield.

3.2.1 Problem statement

We define a bounded region ζ_1 of radius R, as shown in Fig. 3.1 with green-shaded region, such that the desired sound sources are located inside ζ_1. We further as-

sume that the interfering and noise sources as well as reflections originate in a non-overlapping zone ζ_2 (grey-shaded region in Fig. 3.1) and there exist a finite separation between the desired and undesired zones. By measuring the mixed soundfield in the region between ζ_1 and ζ_2, we intend to isolate the overall contribution of the desired and undesired sound sources.

3.2.2 Modal framework

Without loss of generality, the centre of ζ_1 is taken as the global origin. For a height-invariant sound propagation, a cylindrical harmonic representation of the soundfield is more appropriate to attain mathematical tractability. In terms of cylindrical harmonics, a soundfield observed at $\boldsymbol{x} \equiv (r, \phi)$ due to a sound source at $\boldsymbol{y}_s \equiv (r_s, \phi_s)$ is given by [1, ch. 4]

$$P(\boldsymbol{x}, k) = \begin{cases} \displaystyle\sum_{n=-\infty}^{\infty} \alpha_n(k)\ J_n(kr)\ e^{in\phi}, & r_s > r \quad (3.1) \\ \displaystyle\sum_{n=-\infty}^{\infty} \beta_n(k)\ H_n(kr)\ e^{in\phi}, & r_s < r \quad (3.2) \end{cases}$$

where $k = \frac{2\pi f}{c}$ is the *wavenumber*, f denotes frequency, c stands for the speed of sound propagation, $J_n(\cdot)$ and $H_n(\cdot)$ are the n^{th} order Bessel and Hankel functions of first kind, respectively, and $i = \sqrt{-1}$. The n^{th} order cylindrical harmonic coefficients $\alpha_n(k)$ and $\beta_n(k)$ represent the interior and exterior soundfields, respectively. Note that, cylindrical harmonic expansion of a height-invariant soundfield does not contain the height-dependent coordinate z, hence, z is omitted in the definition of $\boldsymbol{x}$ and $\boldsymbol{y}_s$. Furthermore, for a unit amplitude sound source at $\boldsymbol{y}_s$, $\alpha_n(k)$ and $\beta_n(k)$ are defined as [61]

$$\alpha_n(k) = H_n(kr_s)\ e^{-in\phi_s} \quad (3.3)$$

$$\beta_n(k) = J_n(kr_s)\ e^{-in\phi_s}. \quad (3.4)$$

For L desired sound sources in ζ_1 and M interfering sources in ζ_2, total observed sound pressure at $\boldsymbol{x}$ is

$$P(\boldsymbol{x}, k) = \sum_{m=1}^{M} P_m(\boldsymbol{x}, k) + \sum_{\ell=1}^{L} P_\ell(\boldsymbol{x}, k) \tag{3.5}$$

where $P_m(\boldsymbol{x}, k)$ and $P_\ell(\boldsymbol{x}, k)$ are the sound pressures due to m^{th} and l^{th} source, respectively. Note that, the reflections originating from outside ζ_1 can be modelled as image sources [190] and hence, are included in the count of L.

By constraining $\boldsymbol{x}$ to lie between ζ_1 and ζ_2, M sources inside ζ_1 and L sources from ζ_2 contribute to an exterior and interior soundfield at $\boldsymbol{x}$, respectively. Hence, we express (3.5) in the modal domain by comparing it with (3.1) and (3.2):

$$P(\boldsymbol{x}, k) = \sum_{n=-\infty}^{\infty} \left[\left(\sum_{m=1}^{M} \beta_n^{(m)}(k) S_m(k) \right) H_n(kr) + \left(\sum_{\ell=1}^{L} \alpha_n^{(\ell)}(k) S_\ell(k) \right) J_n(kr) \right] e^{in\phi} \tag{3.6}$$

$$= \sum_{n=-\infty}^{\infty} \Big[\beta_n(k) H_n(kr) + \alpha_n(k) J_n(kr) \Big] e^{in\phi} \tag{3.7}$$

where $\alpha_n^{(\ell)}(k)$ and $\beta_n^{(m)}(k)$ are the interior field coefficients due to ℓ^{th} source and exterior field coefficients due to m^{th} source, respectively, whereas $\alpha_n(k)$ and $\beta_n(k)$ are the combined contribution from the corresponding zones ζ_1 and ζ_2. The individual source strengths $S_\ell(k)$ and $S_m(k)$, respectively for ℓ^{th} and m^{th} source, are shown for brevity, however, at this time we do not seek to extract individual sources. Instead, we concentrate on separating the interior and exterior soundfields as a whole by measuring the mixed sound pressure $P(\boldsymbol{x}, k)$ at different spatial points inside the measuring zone. We intend to estimate $\alpha_n(k)$ and $\beta_n(k)$ in order to reconstruct soundfields caused by sound sources originated from the desired or undesired sound zone. It is worth mentioning that the estimation of the desired soundfield is important for audio capturing and reproduction whereas an accurate approximation of the undesired soundfield is a prerequisite for applications like spatial active noise cancellation.

3.2.3 Extracting soundfield coefficients

To separate the interior and exterior soundfield coefficients from the mixed observation, we employ the dual surface approach of near-field acoustical holography [1]. We place two concentric circular microphone arrays with respective radius of r_1 and r_2 in the measuring zone such that $r_2 > r_1 > R$ (Fig. 3.1). Hence, based on the modal expression of a soundfield as in (3.7), the measured sound pressures at the two microphone arrays are given by

$$P(\boldsymbol{x}_1, k) = \sum_{n=-\infty}^{\infty} \Big[\beta_n(k) H_n(kr_1) + \alpha_n(k) J_n(kr_1)\Big] e^{in\phi} \tag{3.8}$$

$$P(\boldsymbol{x}_2, k) = \sum_{n=-\infty}^{\infty} \Big[\beta_n(k) H_n(kr_2) + \alpha_n(k) J_n(kr_2)\Big] e^{in\phi} \tag{3.9}$$

where $\boldsymbol{x}_1 \equiv (r_1, \phi)$ and $\boldsymbol{x}_2 \equiv (r_2, \phi)$. Multiplying both the sides of (3.8) by $e^{-in'\phi}$ and integrating with respect to ϕ, we get

$$\int_0^{2\pi} P(\boldsymbol{x}_1, k)\; e^{-in'\phi}\; d\phi = \sum_{n=-\infty}^{\infty} \Big[\beta_n(k) H_n(kr_1) + \alpha_n(k) J_n(kr_1)\Big] \int_0^{2\pi} e^{i(n-n')\phi}\; d\phi. \tag{3.10}$$

The exponential functions obey the following orthogonality property

$$\int_0^{2\pi} e^{i(n-n')\phi}\; d\phi = \begin{cases} 2\pi, & \text{if } n = n' \\ 0, & \text{otherwise.} \end{cases} \tag{3.11}$$

Using (3.11) in (3.10), we obtain

$$\beta_n(k) H_n(kr_1) + \alpha_n(k) J_n(kr_1) = P_{1n}(k) \tag{3.12}$$

where

$$P_{1n}(k) \equiv \frac{1}{2\pi} \int_0^{2\pi} P(\boldsymbol{x}_1, k)\; e^{-in\phi}\, \mathrm{d}\phi. \tag{3.13}$$

Following the similar technique, we get the following identity from (3.9) and (3.11)

$$\beta_n(k) H_n(kr_2) + \alpha_n(k) J_n(kr_2) = P_{2n}(k) \tag{3.14}$$

where

$$P_{2n}(k) \equiv \frac{1}{2\pi}\int_0^{2\pi} P(\boldsymbol{x}_2, k)e^{-in\phi}\,\mathrm{d}\phi. \tag{3.15}$$

The realisation of (3.13) and (3.15) imposes an impractical requirement of continuous microphone arrays. Hence, they need to be approximated based on a finite number of microphones by virtue of a spatial sampling theory [71] to estimate $P_{1n}(k)$ and $P_{2n}(k)$ as

$$\widehat{P}_{1n}(k) = \frac{1}{2\pi}\sum_{q=1}^{Q_1} A_{1q}P_1(\phi_q, k)e^{-in\phi_q} \tag{3.16}$$

$$\widehat{P}_{2n}(k) = \frac{1}{2\pi}\sum_{q=1}^{Q_2} A_{2q}P_2(\phi_q, k)e^{-in\phi_q} \tag{3.17}$$

where $Q_j\ \forall j \in [1,2]$ is the number of microphones in j^{th} array and ϕ_q denotes the angular location of q^{th} microphone. We restrict Q_j to its minimum value that guarantees to evade spatial aliasing [71]

$$Q_j = (2N_j + 1),\ \forall j \in [1,2] \tag{3.18}$$

where $N_j = \lceil ker_j/2 \rceil$ is the truncated soundfield order [60] and $\lceil\cdot\rceil$ denotes ceiling operation. A_{1q} and A_{2q} are corresponding microphone weights to ensure the validity of the orthonormality property of exponential functions with a finite number of elements. For mathematical tractability, we use a uniform-angle sampling scheme with $A_{1q} = 2\pi/Q_1$ and $A_{2q} = 2\pi/Q_2$. Consequently, we solve (3.12) and (3.14) for $\alpha_n(k)$ and $\beta_n(k)$ as

$$\widehat{\alpha}_n(k) = \frac{H_n(kr_2)\widehat{P}_{1n}(k) - H_n(kr_1)\widehat{P}_{2n}(k)}{J_n(kr_1)H_n(kr_2) - J_n(kr_2)H_n(kr_1)} \tag{3.19}$$

$$\widehat{\beta}_n(k) = \frac{J_n(kr_1)\widehat{P}_{2n}(k) - J_n(kr_2)\widehat{P}_{1n}(k)}{J_n(kr_1)H_n(kr_2) - J_n(kr_2)H_n(kr_1)}. \tag{3.20}$$

In this work, we focus on estimating $\widehat{\beta}_n(k)$ in order to extract exterior soundfield. In the subsequent sections, we demonstrate a few practical applications and experimental validation of the proposed technique based on exterior soundfield estima-

tion. Note that the proposed approach is equally pertinent for interior soundfield extraction for applications such as active noise cancellation.

3.2.4 Practical applications

A couple of practical applications of exterior soundfield extraction are presented in this section.

Selective capturing of a soundfield

The proposed method is useful in capturing soundfields originated from a specific region. Assuming all the undesired sources lie outside the target zone, the desired soundfield can be reconstructed by using (3.20) in (3.2):

$$\widehat{P}_d(\boldsymbol{x}, k) = \sum_{n=-N_d}^{N_d} \widehat{\beta}_n(k)\ H_n(kr)\ e^{in\phi} \tag{3.21}$$

where $\widehat{P}_d(\boldsymbol{x}, k)$ is the soundfield measured at point $\boldsymbol{x}$ caused by the desired sources and $N_d = \lceil ker_d/2 \rceil$ is the exterior soundfield order. r_d is the radius of the farthest source from origin which needs to be known as *a priori* or estimated using a suitable localisation algorithm.

Speech dereverberation

The reflected waves in a reverberant room form an interior soundfield irrespective of the source and microphone positions. Hence, it is possible to model the room reflections based on the image source method [190]. Under such an assumption, (3.6) can be used to represent a reverberant room by letting $M = 1$, $S_m(k) = S(k)$ being the direct path signal, and L as the total number of reflections and noise sources with $S_\ell(k)$ containing corresponding amplitude and phase. Under such a condition, (3.21) represents the spatial filtered version of $S(k)$, i.e., the direct path signal measured at $\boldsymbol{x}$. Furthermore, with the knowledge of source DOA, we can

use (3.4), (3.6) and (3.21) to estimate unfiltered version of $S(k)$ by

$$\widehat{S}(k) = \frac{\widehat{P}_s(\boldsymbol{x}, k)}{\sum_{n=-N_S}^{N_S} \beta_n(k) H_n(kr) e^{in\phi}}. \tag{3.22}$$

Note that, $S(k)$ can also be estimated solely based on a single mode by $\widehat{S}(k) = \widehat{\beta}_n/\beta_n \; \forall n$, however, (3.22) is the preferable approach to avoid occasional Bessel zeros.

3.3 3D sound Propagation Model

A 3D sound propagation is more generic and practical model compared to the height-invariant case. The model we developed in the last section to separate interior and exterior soundfield for height-invariant sound propagation is conceptually applicable for 3D soundfield as well, however, it requires a large number of microphones for a large 3D region. Hence, in this section, we devise an alternative approach to solve the sound separation problem for a 3D soundfield.

3.3.1 Problem description

The spherical harmonic decomposition of a 3D soundfield at any point $\boldsymbol{x} \equiv (r, \theta, \phi)$ due to a sound source at $\boldsymbol{y}_s \equiv (r_s, \theta_s, \phi_s)$ is given by [1, ch. 6]

$$P_I(\boldsymbol{x}, k) = \sum_{nm}^{N_I} \alpha_{nm}(k) \; j_n(kr) \; Y_{nm}(\widehat{\boldsymbol{x}}), \quad \text{if } r_s > r \tag{3.23}$$

$$P_E(\boldsymbol{x}, k) = \sum_{nm}^{N_E} \beta_{nm}(k) \; h_n(kr) \; Y_{nm}(\widehat{\boldsymbol{x}}), \quad \text{if } r_s < r \tag{3.24}$$

where $\alpha_{nm}(k)$ and $\beta_{nm}(k)$ are the spherical harmonic coefficients for interior soundfield P_I and exterior soundfield P_E, respectively, $Y_{nm}(\widehat{\boldsymbol{x}})$ is the spherical harmonic of order n and degree m towards $\widehat{\boldsymbol{x}} \equiv (\theta, \phi)$, and $j_n(\cdot)$ and $h_n(\cdot)$ are the n^{th} order spherical Bessel and Hankel functions of the first kind, respectively. The truncation limits of the soundfield orders are given by $N_I = \lceil ker/2 \rceil$ and $N_E = \lceil ker_s/2 \rceil$

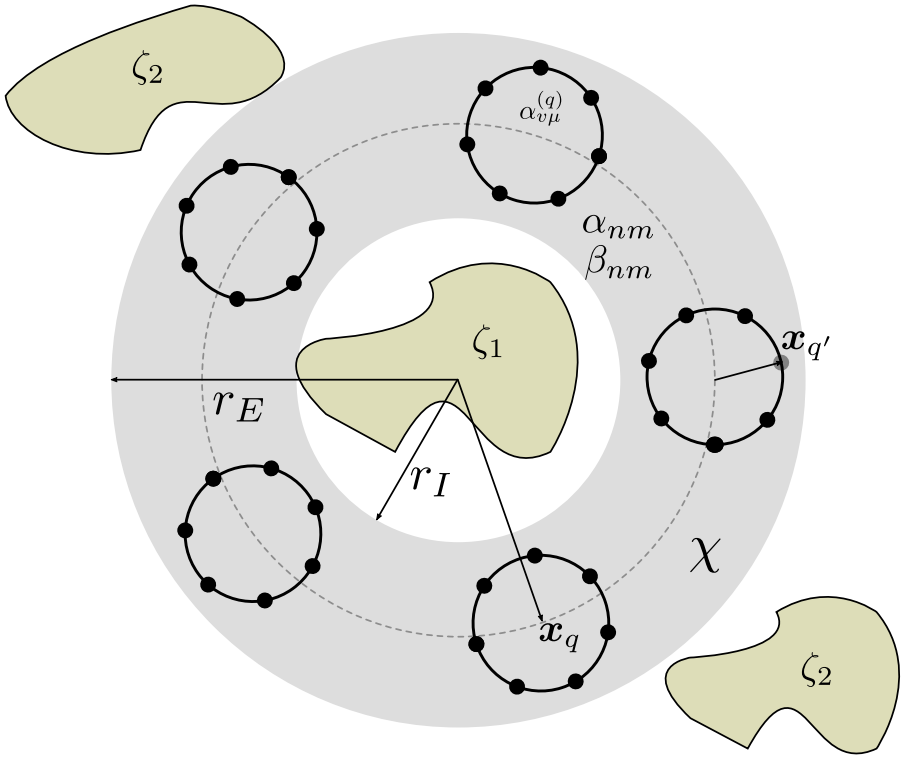


Figure 3.2: Projection of a 3D soundfield separation aperture setup.

[60]. For a 3D soundfield, we define a spherical shell region χ as the measuring zone with an inner radius of r_I and outer radius r_E (Fig. 3.2). Restricting all the desired sources ζ_1 and undesired sources ζ_2 inside and outside the spherical shell, respectively, the observed sound pressure inside χ follows the superposition principle, i.e.,

$$P(\boldsymbol{x}, k) = P_E(\boldsymbol{x}, k) + P_I(\boldsymbol{x}, k) \tag{3.25}$$

where $P_E(\boldsymbol{x}, k)$ and $P_I(\boldsymbol{x}, k)$ are the contributions of the desired and undesired sources, respectively. Given the measured sound pressure $P(\boldsymbol{x}, k)$, our goal is to estimate $\alpha_{nm}(k)$ and $\beta_{nm}(k)$ to separate interior and exterior soundfields using (3.23) and (3.24).

3.3.2 Soundfield separation using an array of HOMs

A higher order microphone (HOM) itself is an array of pressure microphones which is capable of recording higher order soundfields. Let us consider an array of Q

HOMs of order V. Each HOM located at a position $\boldsymbol{x}_q \equiv (r_q, \theta_q, \phi_q)$, where $q \in [1, Q]$. We further assume that each HOM contains Q' pressure microphones located at $\boldsymbol{x}_{q'} \equiv (r_M, \theta_{q'}, \phi_{q'})$ with respect to the local origin of the corresponding HOM, where $q' \in [1, Q']$. Note that all the desired and undesired sources create an interior soundfield on each HOM irrespective of their positions. Hence, the local soundfield coefficient for each HOM in the array is given by (2.41)

$$\alpha_{v\mu}^{(q)} = \frac{1}{b_v(kr_M)} \sum_{q'=1}^{Q'} P(\boldsymbol{x}_{q'}, k)\ Y_{v\mu}^{*}(\widehat{\boldsymbol{x}}_{q'}) \tag{3.26}$$

where * denotes complex conjugate, $\widehat{\boldsymbol{x}}_{q'} \equiv (\theta_{q'}, \phi_{q'})$, $v \in [0, V]$ and $\mu \in [-v, v]$ are local order and degree of each HOM, respectively, and

$$b_v(kr_M) = \begin{cases} j_v(kr_M) & \text{for an open array} \\ j_v(kr_M) - \frac{j_v'(kr_M)}{h_v'(kr_M)} h_v(kr_M) & \text{for a rigid array.} \end{cases} \tag{3.27}$$

Applying the addition theorem for Bessel and Hankel functions [62], [191] in (3.23), (3.24), (3.25) and (3.26), we can relate the global coefficients α_{nm} and β_{nm} with the local coefficients $\alpha_{v\mu}^{(q)}$ as

$$\alpha_{v\mu}^{(q)}(k) = \sum_{nm}^{N_I} \alpha_{nm}(k)\ \widehat{S}_{nv}^{m\mu}(\boldsymbol{x}_q) + \sum_{nm}^{N_E} \beta_{nm}(k)\ S_{nv}^{m\mu}(\boldsymbol{x}_q) \tag{3.28}$$

where

$$\widehat{S}_{nv}^{m\mu}(\boldsymbol{x}_q) = 4\pi i^{(v-n)} \sum_{\ell=0}^{n+v} i^{\ell}(-1)^m\ j_\ell(kr_q)\ Y_{\ell(m-\mu)}(\widehat{\boldsymbol{x}}_q)\ W_1 W_2\ \xi \tag{3.29}$$

$$S_{nv}^{m\mu}(\boldsymbol{x}_q) = 4\pi i^{(v-n)} \sum_{\ell=0}^{n+v} i^{\ell}(-1)^m\ h_\ell(kr_q)\ Y_{\ell(m-\mu)}(\widehat{\boldsymbol{x}}_q)\ W_1 W_2\ \xi \tag{3.30}$$

with

$$W_1 = \begin{pmatrix} n & v & \ell \\ 0 & 0 & 0 \end{pmatrix} \text{ and } W_2 = \begin{pmatrix} n & v & \ell \\ -m & \mu & (m-\mu) \end{pmatrix} \tag{3.31}$$

denoting Wigner 3-j symbols [189], $\widehat{\boldsymbol{x}}_q \equiv (\theta_q, \phi_q)$, and

$$\xi = \sqrt{\frac{(2n+1)(2v+1)(2\ell+1)}{4\pi}}. \tag{3.32}$$

Rewriting (3.28) in a matrix form, we get

$$\boldsymbol{\alpha} = \boldsymbol{T}\boldsymbol{d} \tag{3.33}$$

where

$$\boldsymbol{\alpha} \in \mathbb{C}^{Q(V+1)^2} = [\alpha_{00}^{(1)}(k), \ldots \alpha_{VV}^{(1)}(k), \ldots \alpha_{00}^{(Q)}(k), \ldots \alpha_{VV}^{(Q)}(k)]^T \tag{3.34}$$

$$\boldsymbol{d} \in \mathbb{C}^{(N_I+1)^2+(N_E+1)^2} = [\alpha_{00}(k), \ldots \alpha_{N_I N_I}(k), \beta_{00}(k), \ldots \beta_{N_E N_E}(k)]^T \tag{3.35}$$

$$\boldsymbol{T} = \begin{bmatrix} \widehat{S}_{00}^{00}(\boldsymbol{x}_1) & \ldots & \ldots & \widehat{S}_{N_I 0}^{N_I 0}(\boldsymbol{x}_1) & S_{00}^{00}(\boldsymbol{x}_1) & \ldots & \ldots & S_{N_E 0}^{N_E 0}(\boldsymbol{x}_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{S}_{0V}^{0V}(\boldsymbol{x}_Q) & \ldots & \ldots & \widehat{S}_{N_I V}^{N_I V}(\boldsymbol{x}_Q) & S_{0V}^{0V}(\boldsymbol{x}_Q) & \ldots & \ldots & S_{N_E V}^{N_E V}(\boldsymbol{x}_Q) \end{bmatrix} \tag{3.36}$$

with $\mathbb{C}$ being the set of complex numbers. Eq. (3.33) can be solved for the soundfield coefficients $\boldsymbol{d}$ by

$$\widehat{\boldsymbol{d}} = \boldsymbol{T}^{\dagger}\boldsymbol{\alpha} \tag{3.37}$$

where $(\cdot)^{\dagger}$ denotes pseudo-inverse operation on a matrix. Note that, to avoid an under-determined system of (3.33), at least $Q_{\min}$ HOMs are required in the array, where

$$Q_{\min} = \frac{(N_I+1)^2 + (N_E+1)^2}{(V+1)^2}. \tag{3.38}$$

3.3.3 Practical applications

Similar to the height-invariant case, the estimated α_{nm} and β_{nm} can be used in reconstructing desired soundfield. A couple of examples are given below.

Interior-exterior separation

We attain interior and exterior separation of a mixed soundfield by using $\widehat{\alpha}_{nm}(k)$ and $\widehat{\beta}_{nm}(k)$ from (3.37) in (3.23) and (3.24) as

$$\widehat{P}_I(\boldsymbol{x}, k) = \sum_{nm}^{N_I} \widehat{\alpha}_{nm}(k)\ j_n(kr)\ Y_{nm}(\widehat{\boldsymbol{x}}) \tag{3.39}$$

$$\widehat{P}_E(\boldsymbol{x}, k) = \sum_{nm}^{N_E} \widehat{\beta}_{nm}(k)\ h_n(kr)\ Y_{nm}(\widehat{\boldsymbol{x}}). \tag{3.40}$$

Equation (3.40) eliminates all the undesired sources outside r_E along with external noise and reflections. Conversely, (3.39) is useful in estimating undesired soundfield to compute the driving signal for the secondary speakers in an active noise cancellation task.

Sound source extraction

$\widehat{P}_E(\boldsymbol{x}, k)$ of (3.40) can be considered as the free field response at $\boldsymbol{x}$ due to a point source excitation $S_E(k)$ at $\boldsymbol{y}_s \equiv (r_s \leq r_E, \theta_s, \phi_s)$. Hence, the source excitation can be estimated as

$$\widehat{S}_E(k) = \frac{\widehat{P}_E(\boldsymbol{x}, k)}{P_E(\boldsymbol{x}, k)} \tag{3.41}$$

where the corresponding free-field unit amplitude response $P_E(\boldsymbol{x}, k)$ is obtained by using the analytical value of $\beta_{nm}(k)$ from (2.26) in (3.24).

3.4 Experimental Results

This section contains experimental evaluations and analysis of the performances of the two algorithms we have discussed in this chapter.

3.4.1 Dual surface approach for height-invariant soundfield

First we demonstrate the performance of the dual surface approach using a height-invariant sound propagation model. Unless specified otherwise, the following pa-

rameter settings were used for the simulations: $r_1 = 1$ m, $r_2 = 1.5$ m, $r_s = 0.5$ m, $\phi_s = \pi/3$ and $c = 343$ m/s.

Array design

Theoretically, for a perfect reconstruction, we require at least $Q_1 = (2N_1 + 1)$ and $Q_2 = (2N_2+1)$ microphones in array 1 and array 2, respectively. At high frequency, it requires a large number of microphones to achieve aliasing-free reproduction of a soundfield. However, in some practical scenarios, it may deem acceptable and more appropriate to reduce the number of microphones at the expanse of constrained reproduction error. In this exercise, we gradually introduced spatial aliasing error by reducing number of microphones in the arrays and observed its impact in spatial reconstruction. We are going to show that for a bounded aliasing error, the estimation error remains within an acceptable limit. Note that, the term *acceptable* is subjective, and depends on the usage and error sensitivity of the system.

For simplicity, we used the same number of microphones Q for both the arrays in our simulation. To measure the performance, we defined the following coefficient estimation error

$$C_{\text{err}} = \frac{\sum_{\forall n} |\beta_n(k) - \widehat{\beta}_n(k)|}{\sum_{\forall n} |\beta_n(k)|} \tag{3.42}$$

where $|\cdot|$ denotes absolute value. We used 20 random sources, 10 point sources and 10 plane waves, to simulate the undesired soundfield using the 2D wave propagation model as [186]

$$P_\ell(\boldsymbol{x}, k) = \begin{cases} A_\ell \ \frac{i}{4} \ H_0(k \, \|\boldsymbol{x} - \boldsymbol{y}_\ell\|) & \text{for point source,} \\ A_\ell \ e^{-ik \ \widehat{\boldsymbol{y}}_\ell \cdot \boldsymbol{x}} & \text{for plane wave} \end{cases} \tag{3.43}$$

where $\|\cdot\|$ denotes Euclidean distance. The signal magnitude A_ℓ and source location $\boldsymbol{y}_\ell$ were chosen randomly outside r_2.

Fig. 3.3 plots C_{err} against various Q in distinct frequencies. As expected, C_{err} is negligible when $Q \geq Q_2$. However, we can make an intriguing observation in the region $Q_1 \leq Q \leq Q_2$ where C_{err} remains low up to some extent. This is likely due

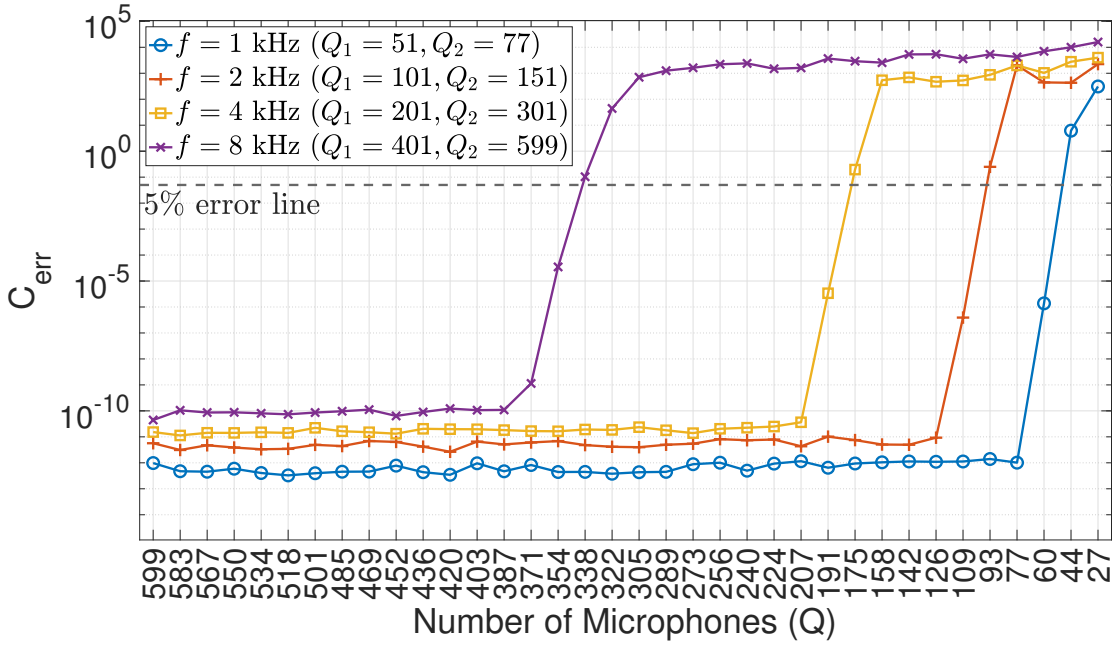


Figure 3.3: $\beta_n(k)$ estimation error for different number of microphones in the arrays.

to the fact that, the exterior soundfield order is determined by the source radius R (Fig. 3.1), thus the exterior soundfield order remains bounded by $N_s < N_1 < N2$ in the proposed model. Hence, the contribution of the higher orders gradually decreases beyond N_1 for the exterior soundfield. We can further observe that, as we start reducing the number of microphones (Q) in the array from the theoretical lower bound of Q_2, the impact of spatial aliasing on estimation error becomes prominent much earlier in the lower frequencies. Therefore, depending on the frequency, array radii, and microphone spacing, it is possible to capture an exterior soundfield with $Q < Q_2$. However, for estimating an interior soundfield to capture the undesired sources, it is required to obey the theoretical limit of $Q \geq Q_2$. In this work, we chose Q in heuristic manner based on the highest frequency component of the signal.

Exterior soundfield recording

The next set of simulations was designed to extract desired soundfield from the same mixed acoustic scenario we used in the previous section. The environment was set to have $f = 1$ kHz, $L = 20$, $M = 1$, and $Q = 70$ which was intentionally kept

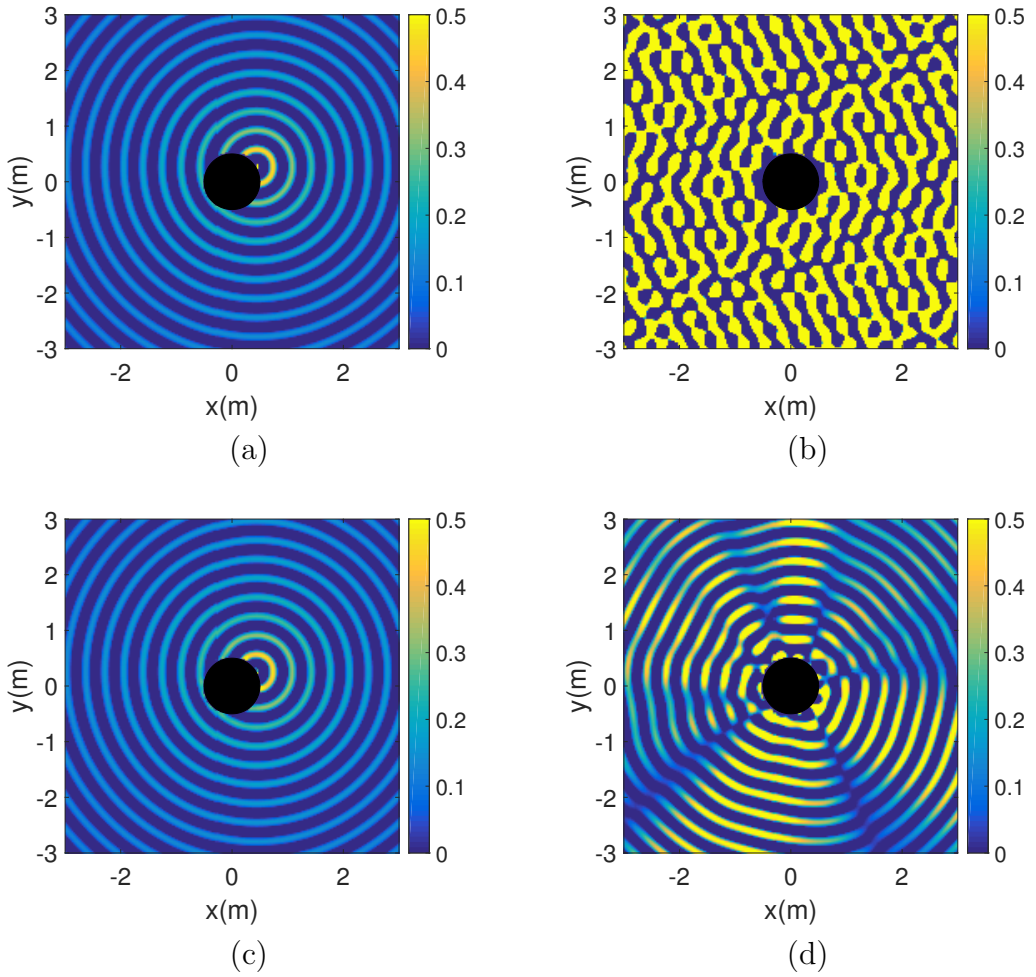


Figure 3.4: soundfield recording with $f = 1$ kHz, $Q = 70$, $r_1 = 1$ m, $r_2 = 1.5$ m, $r_s = 0.5$ m and $\phi_s = \pi/3$. The black circle denotes the source area. (a) Original soundfield, (b) combined soundfield with 20 undesired sources outside r_2, (c) reconstructed soundfield without thermal noise and (d) reconstructed soundfield with 30 dB thermal noise.

lower than Q_2. Fig. 3.4(a) and 3.4(b) show the desired and the mixed soundfields, respectively. We performed the extraction process under two different conditions. In the first case, we explored the ideal case with no measurement inaccuracy which resulted in a perfect reconstruction of the desired soundfield (Fig. 3.4(c)).

In real world applications, often the measurements we obtain get tainted by the external factors and deviations. One of the major reasons for inaccurate measure-

ments is the thermal noise that exists at the pressure sensors of the microphones caused by the thermal vibrations. This induces random noise in the observed microphone signals, P_1 and P_2 of (3.8) and (3.9), respectively. To replicate a realistic scenario and evaluate the performance of the proposed algorithm under practical deviations, we added random white Gaussian noise at each microphone while maintaining a combined signal to noise ratio (SNR) of 30 dB at each microphone array. The SNR at each microphone array was calculated by

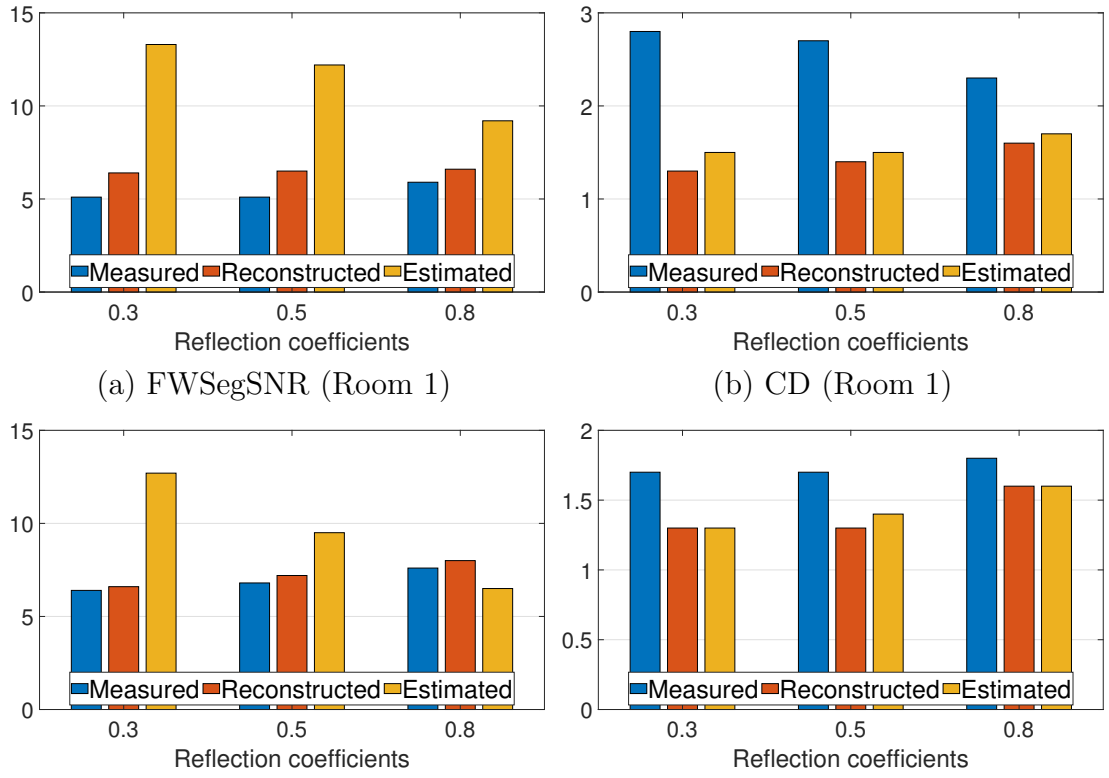
$$\mathrm{SNR}_j = \frac{\sum_{q=1}^{Q_j} |P_j(\phi_q, k)|^2}{\sigma^2 Q_j}, \text{ for } j \in [1, 2] \tag{3.44}$$

where σ^2 is the noise power.

The estimated soundfield in the presence of thermal noise is shown in Fig. 3.4(d). The thermal noise introduces a low level of distortion in the expected outcome, however, it still exhibits a good resemblance with the intended soundfield. Note that, the distortion is entirely due to the measurement inaccuracies caused by the microphone internal noise, as the proposed method inherently cancels out all the external noise sources lying outside r_2. Furthermore, the seemingly directional energy flux is random and changes its direction in each realisation.

Application in broadband soundfield separation

So far, we have measured the performance of the proposed algorithm using narrowband signals. However, as many acoustic scenarios involve broadband processing, this section demonstrates the proposed algorithm's ability to perform the soundfield separation of broadband signals. To this end, we apply the proposed technique to achieve speech dereverberation in different reverberant environments simulated using the image source method [190]. The image source method imitates the reflections as point sources at varying distances. Hence, in the context of this work, the reflections can be considered originating from external point sources which contribute to interior soundfield in the microphone positions. Our objective is to suppress the reflections in order to extract the direct path signal by virtue of the proposed interior-exterior soundfield separation technique.



(a) FWSegSNR (Room 1) (b) CD (Room 1)

(c) FWSegSNR (Room 2) (d). CD (Room 2)

Figure 3.5: Dereverberation performance in the presence of 30dB thermal noise. The legend entry *Reconstructed* indicates the direct path signal measured at $(r_1, \pi/3)$ whereas *Estimated* denotes the extraction of the source signal.

We evaluated speech dereverberation performance in two different hypothetical 2D rooms with $[8 \times 8]$ m and $[5 \times 5]$ m dimensions. We used the image method to simulate reverberant conditions in those rooms with (r_1, r_2, r_s) being $(2.5, 3, 1)$ m and $(1, 1.5, 0.5)$ m, respectively. Each image source was modelled as a 2D point source in space. Clean speech data were taken from WSJCAM0 corpus [192] and re-sampled at $f_s = 8$ kHz to reduce computational cost. A 256-point discrete Fourier transform was used with a 20 ms window and 50% overlap. We evaluated the performance of dereverberation in terms of frequency-weighted segmental SNR (FWSegSNR) and cepstral distance (CD) [193]. For reference, an increase in FWSegSNR and reduction in CD indicate a better performance in terms of dereverberation and background noise suppression.

For a Nyquist frequency of 4 kHz, the required numbers of microphone are given by (3.18) as $[Q_1, Q_2] = [499, 599]$ and $[201, 301]$ for room 1 and 2, respectively. However, based on the discussion on array design in the previous section, we used a reduced number of microphones $Q = 325$ and 175 for room 1 and 2, respectively. The simulation results are shown in Fig. 3.5 for different reflection coefficients. We measured the performance of reconstructed exterior soundfield $\widehat{P}$ and estimated clean speech $\widehat{S}$ at $(r_1, \pi/3)$ in the presence of thermal noise. We observe up to 1 dB and 1.5 dB improvements in terms of FWSegSNR and CD for $\widehat{P}$, respectively. However, a significant improvement of FWSegSNR is detected in case of $\widehat{S}$, except for room 2 at high reverberant condition which can be caused by the truncation and aliasing error. As FWSegSNR measures the spectral similarities, $\widehat{S}$ shown a better result compared to the spatially filtered version of $\widehat{P}$. The improved SNR for $\widehat{S}$ comes at the cost of an additional requirement to know the source position contrary to $\widehat{P}$ which can be estimated only from the knowledge of source radius. The latter case is particularly useful when the speakers are located at fixed radial positions (e.g., in a meeting room). Note that, CD between the processed output does not exhibit a large difference which indicates that both $\widehat{S}$ and $\widehat{P}$ attain the same level of performance in terms of speech distortion.

3.4.2 HOM-based approach for 3D sound propagation

In this section, we continue our discussion on performance evaluation of the proposed 3D soundfield separation technique employing an array of HOMs.

Array geometry & simulation criteria

The conventional sampling schemes [73] require to place the microphones around a sphere in a regular pattern, which is often inconvenient for physical implementation. Hence, we propose to distribute the HOMs in 3 distinct planes at 45°, 90° and 135° elevations on the surface of a sphere (Fig. 3.6) and solve the problem in a *least-square* sense to avoid distortion due to irregular distribution of microphones. In the simulation, the radial distance of each HOM was randomised within a spherical shell between $0.8 - 1.0$m to improve robustness against ill-posed problem. We embraced the theory of over-sampling [61] by a factor of $\kappa = 0.75$ and calculated the total

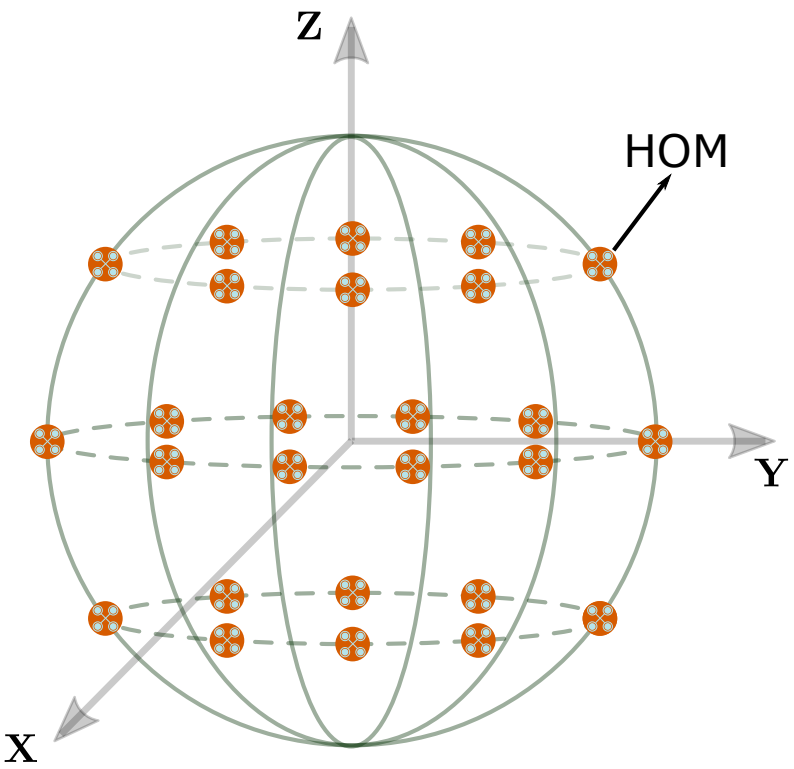
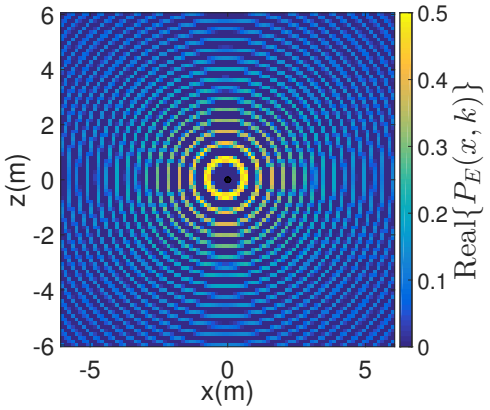
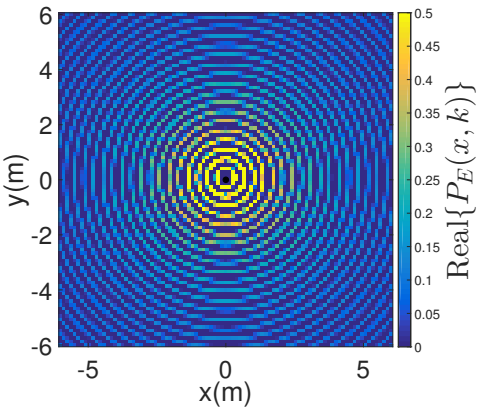


Figure 3.6: The higher order microphones are placed along the dotted lines in 3 rings around the sphere. The desired and undesired sources need to be inside and outside the sphere, respectively.

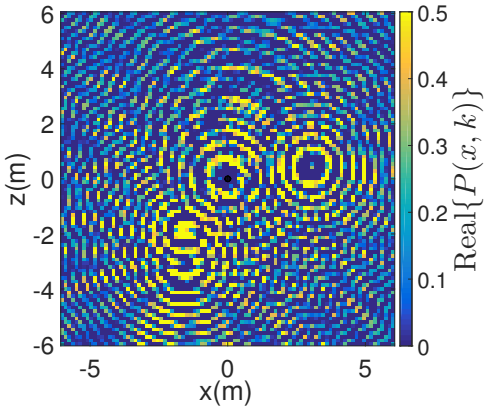
number of HOMs in the array as $Q = Q_{\min}/\kappa$ (i.e. $Q = 39$ at 1 KHz frequency with a source radius of 0.1 m). The HOMs were evenly distributed on the three elevation planes with each plane consisting of $Q/3$ HOMs. Each HOM was assumed to be rigid and of 4^{th} order, i.e. $V = 4$, with uniformly distributed sensors. The sound pressure at each sensor of the HOMs were simulated assuming point sources, however, the method is equally applicable for directional or non-point sources as well. We simulated the thermal noise by adding complex white Gaussian noise at each sensor of the HOM in such a way that the combined array signal to noise ratio remains at 30 dB. All the measurements and calculations were performed in the frequency domain.



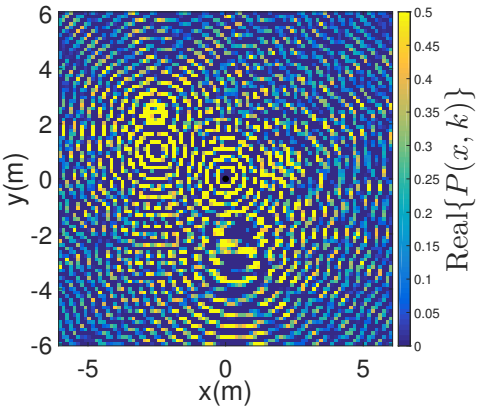
(a) Reference (XZ plane)



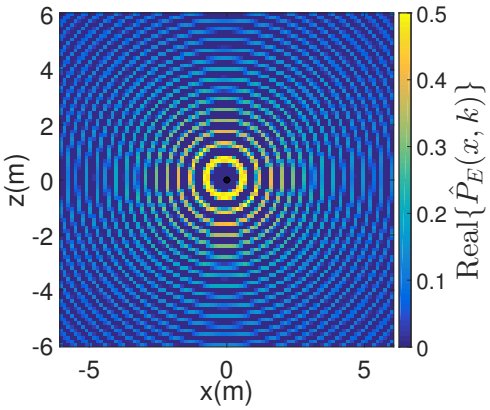
(b) Reference (XY plane)



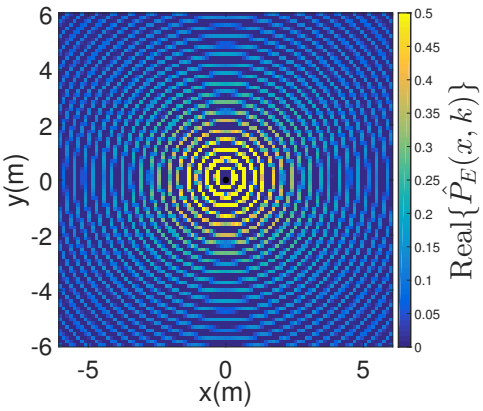
(c) Measured (XZ plane)



(d) Measured (XY plane)



(e) Reconstructed (XZ plane)



(f) Reconstructed (XY plane)

Figure 3.7: The reconstructed soundfields on 2 different planes for a desired source at 0.1m from the origin with 5 random interfering sources. 30 dB thermal noise was added to the microphone measurements.

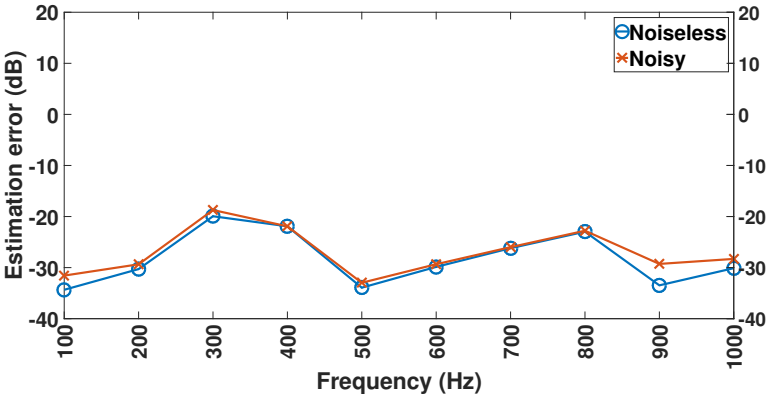
Extraction of an exterior soundfield

Based on the aforementioned setup, we intend to extract the desired soundfield originated from a bounded region in the presence of external interfering sources as well as thermal noise at microphones. The simulations were performed at 1 KHz frequency with $Q = 39$ and a desired source 0.1 m from the origin. We considered 5 interfering sources distributed randomly outside the spherical shell χ. Fig. 3.7 demonstrates the reconstruction accuracy on 2 different planes. The left column of Fig. 3.7 exhibit, respectively from top to bottom, the reference, measured, and reconstructed soundfield on XZ-plane $y = 1.2$ m. The similar plots are shown on the right-sided column of Fig. 3.7, but with respect to the XY plane at $z = 0$m. The reconstructed soundfields in both the planes confirm that the 3D soundfield separation technique is capable of accurately extracting the exterior soundfields. We also observe that, unlike the height-invariant case in Fig. 3.4, the HOM-based separation does not exhibit any visible distortion due to the measurement inaccuracy contributed by the thermal noise. This certifies that an array of rigid HOMs is more robust in a practical environment compared to the dual surface approach.

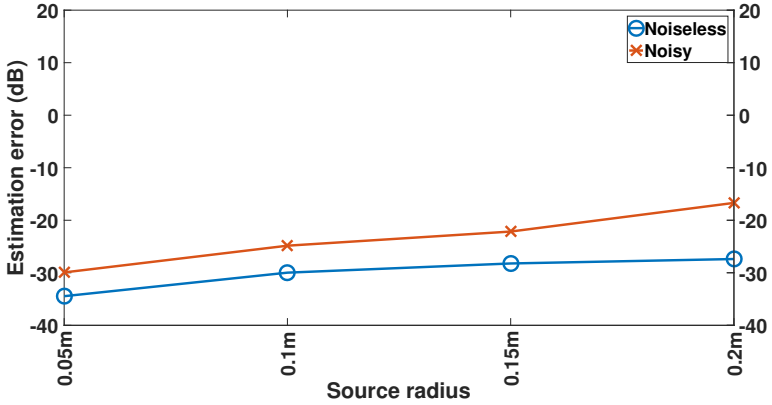
We also measured the estimation error for various source radii and frequencies. The estimation error was defined by

$$\epsilon(k) = \frac{\sum\limits_{\forall \boldsymbol{x}} |P_E(\boldsymbol{x}, k) - \widehat{P}_E(\boldsymbol{x}, k)|^2}{\sum\limits_{\forall \boldsymbol{x}} |P_E(\boldsymbol{x}, k)|^2}. \tag{3.45}$$

The results were accumulated over 10 trials where the desired and interfering sources were positioned randomly in each trial. We considered a 3D region from -6 m to 6 m in each dimension with respect to the origin and calculated average estimation error within the region. Fig. 3.8(a) shows that the estimation error remains low in the designated frequency range. The estimation error also remains insignificant in the plotted radii range, as shown in Fig. 3.8(b), however, it shows a gradually increasing trend for a larger source region. This suggests that as we keep increasing the source radius, the estimation may fail at some point due to mathematical instability of the translation matrix. To overcome this issue, a proper



(a)



(b)

Figure 3.8: Estimation error against different (a) frequencies ($r_E = 0.1$m) and (b) source radii ($f = 1$ KHz). The term "noisy" denotes the presence of measurement inaccuracy due to thermal noise.

regularisation [44] method should be accompanied to solve an ill-posed problem. Note that, the angular source position does not have any significant impact on the estimation error due to the fact that the soundfield order N_E depends only on the source radius.

Real-world applications

The proposed method offers a proficient conceptual model for 3D soundfield separation in a mixed acoustic environment. The algorithm offers numerous application in real-world scenarios as we discussed in the preceding chapters. However, it re-

quires a large number of microphones to deal with a large spatial region and high frequency contents. Hence, at the current state, the concept is more practical to apply in certain areas that deal with low frequencies, such as active noise cancellation, or a smaller spatial region like a wearable recording device in a aircraft cockpit. Furthermore, it is possible to implement the idea with a reduced number of microphones by imposing certain constraints or assuming priors on the acoustic environments as done in the following subsequent works [194], [195].

3.5 Summary

This work introduced soundfield separation techniques for 2D and 3D soundfields based on near-field acoustical holography. The followings are the major contributions and findings reached at the end of this work:

- A soundfield separation technique was devised for height-invariant sound propagation using two concentric circular arrays. The theoretical development was complemented by experimental validation in terms of soundfield reproduction and speech dereverberation under various acoustic conditions.
- We developed a novel technique of 3D soundfield separation by applying near-field acoustical holography with higher order microphones. The advantages of employing higher order microphones in near-field acoustical holography are multi-folds: (1) it allows soundfield separation using a single holographic plane, (2) it offers a logistical advantage by reducing the microphone density, (3) it improves robustness of system due to the inherent characteristics of rigid HOMs.
- We showed that both the techniques are capable of extracting exterior soundfields from mixed measurements irrespective of the number and nature of interfering sources.
- The HOM-based model was found to be more robust against thermal noise at the microphones.
- We proposed a sparse distribution of higher order microphones to reduce the design complexity by offering a simpler array geometry.

- We performed multiple stress tests with both the models to investigate their robustness against the variation of frequencies, number of microphones, and the size of the source region.

The soundfield separation technique is useful as a standalone application in different fields of acoustics. It can also be used as a pre-processing tool to a source separation algorithm to suppress the undesired soundfields and enhance the performance. In the next chapter, we are going to focus on developing a mathematical model for the modal coherence of a soundfield which can be exploited in accomplishing various signal processing tasks such as source separation and DOA estimation.

3.6 Related Publications

- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "Sound field separation in a mixed acoustic environment using a sparse higher order spherical microphone array", Proc. Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), pp. 151–155, San Francisco, USA, October 2017.

- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "Extraction of exterior field from a mixed sound field for 2D height-invariant sound propagation", Proc. International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 1–5, Xian, China, September 2016.

This page intentionally left blank.

Chapter 4

PSD Estimation from Modal Coherence of a Noisy and Reverberant Soundfield

In the previous chapter, we devised multiple techniques for spatial separation of a soundfield. However, various practical applications require a finer dissection of a soundfield down to its individual source components. In pursuit of an efficient source separation algorithm, this chapter lay the basic foundation by developing a mathematical model for the modal coherence of the spherical harmonic coefficients of a noisy and reverberant soundfield in the presence of multiple sound sources. Subsequently, we exploit the model to estimate the power spectral densities (PSD) of the individual sound components in a *least-square* sense. We also investigate certain implementation issues and offer engineering solutions to them. The performance of the proposed algorithm is evaluated in real-life environments using a commercially available microphone array in order to incorporate the deviations incurred in a real-world scenario.

4.1 Introduction

The spherical harmonic decomposition of a soundfield offers an intuitive description of a soundfield behaviour. Hence, by delving into the underlying spatial structure of a soundfield in terms of the modal coherence of its spherical harmonic coefficients, one can expect to better understand the soundfield characteristics and extract its fundamental elements. In this work, we develop a novel mathematical model for a noisy and reverberant soundfield and exploit the model in estimating power spectral densities of the desired signal as well as of the reverberation and coherent noise components in a multi-source environment.

The power spectral density (PSD) of an audio signal carries useful information about the signal characteristics. Many spectral enhancement techniques, most commonly the Wiener filter and spectral subtraction methods, use the knowledge of the PSD to suppress undesired signal components such as background noise[108], late reverberation [109], [196], or both [111]. Other applications of the knowledge of PSD include computing direct to reverberation energy ratio (DRR) [107] or separating sound sources in a mixed acoustic environment [8]. Most of the existing spectral enhancement techniques focused on estimating PSD components under strict assumptions such as a noiseless, free-field or a single-source scenario [8], [107], [110]–[113], [116]. We intend to release those restrictions and approach the problem by decomposing the soundfield into space using a set of orthogonal basis functions.

Exploiting different properties of the spherical harmonics and relevant functions, we develop a model to express a complex soundfield and use that for various purposes. The orthogonality of the spherical harmonic basis functions ensures a well-posed solution without imposing any additional design criteria [113]. Additionally, in contrast to the conventional beamformer-based methods [8] where only the auto-correlation coefficients of the beamformer output were used, we also incorporate the cross-correlation between the spherical harmonic coefficients in our solution. This latter approach was used in [107], [116] for estimating DRR in a single-source environment. The additional correlation coefficients make the algorithm suitable for separating a large number of sources compared to the conventional techniques. We also carry out detailed theoretical analysis, demonstrate the practical impact,

and offer engineering solutions to various implementation challenges such as the Bessel-zero issue which, if not addressed in a correct way, significantly limits the performance of the system.

The rest of the chapter is organised as follows. Section 4.2 contains the problem statement and defines the objective of the work. In Section 4.3, we develop a mathematical model of modal coherence of a noisy and reverberant environment. We use the modal framework in Section 4.4 to devise a PSD estimation technique. Finally in Section 4.5, we evaluate and compare the performance of the proposed algorithm with other contemporary methods based on objective metrics and graphical aids.

4.2 Problem Formulation

Let us consider a microphone array consisting of Q microphones to capture the soundfield in a noisy reverberant room with L distinct sound sources. The received signal at the q^{th} microphone is given by

$$p(\boldsymbol{x}_q, t) = \sum_{\ell=1}^{L} h_\ell(\boldsymbol{x}_q, t) * s_\ell(t) + z(\boldsymbol{x}_q, t) \tag{4.1}$$

where $q \in [1, Q]$, $\ell \in [1, L]$, $\boldsymbol{x}_q \equiv (r_q, \theta_q, \phi_q)$ denotes the q^{th} microphone position, $h_\ell(\boldsymbol{x}_q, t)$ is the RIR between the ℓ^{th} source and the q^{th} microphone, t is the discrete time index, $*$ denotes the convolution operation, $s_\ell(t)$ is the source excitation for the ℓ^{th} sound source, and $z(\boldsymbol{x}_q, t)$ is the coherent noise[1] at the q^{th} microphone position. The RIR can be decomposed into two parts

$$h_\ell(\boldsymbol{x}_q, t) = h_\ell^{(d)}(\boldsymbol{x}_q, t) + h_\ell^{(r)}(\boldsymbol{x}_q, t) \tag{4.2}$$

where $h_\ell^{(d)}(\boldsymbol{x}_q, t)$ and $h_\ell^{(r)}(\boldsymbol{x}_q, t)$ are the direct and reverberant path components, respectively. Substituting (4.2) into (4.1) and converting into frequency domain

[1]Here the coherent noise refers to the coloured background noise, different from the white thermal noise, which can be originated from any unknown noise source such as room air-conditioning system.

using short-time Fourier transform (STFT), we obtain

$$P(\boldsymbol{x}_q, \tau, k) = \sum_{\ell=1}^{L} S_\ell(\tau, k)\Big(H_\ell^{(d)}(\boldsymbol{x}_q, \tau, k) + H_\ell^{(r)}(\boldsymbol{x}_q, \tau, k) \Big) + Z(\boldsymbol{x}_q, \tau, k) \tag{4.3}$$

where $\{P, S, H, Z\}$ represent the corresponding signals of $\{p, s, h, z\}$ in the STFT domain, τ is the time frame index, $k = 2\pi f/c$, f denotes the frequency, and c is the speed of sound propagation. In the subsequent sections, the time frame index τ is omitted for brevity.

Given the measured sound pressure $p(\boldsymbol{x}_q, t)\ \forall q$, we aim to estimate the individual source PSDs, $\mathbb{E}\Big\{|S_\ell(k)|^2\Big\}\ \forall \ell$, where $\mathbb{E}\{\cdot\}$ represents the expected value over time.

4.3 Modal Framework for PSD Estimation

In this section, we develop a spherical harmonic domain framework to establish the relationship between the soundfield coefficients and the individual PSD components in a multi-source noisy and reverberant environment. We use this model in Section 4.4 to estimate individual PSD components from a mixed recording.

4.3.1 Spatial domain representation of room transfer function

We model the direct and reverberant path of room transfer function (RTF) in the spatial domain as

$$H_\ell^{(d)}(\boldsymbol{x}_q, k) = G_\ell^{(d)}(k)\ e^{ik\ \widehat{\boldsymbol{y}}_\ell \cdot \boldsymbol{x}_q} \tag{4.4}$$

$$H_\ell^{(r)}(\boldsymbol{x}_q, k) = \int_{\widehat{\boldsymbol{y}}} G_\ell^{(r)}(k, \widehat{\boldsymbol{y}})\ e^{ik\ \widehat{\boldsymbol{y}} \cdot \boldsymbol{x}_q}\ d\widehat{\boldsymbol{y}} \tag{4.5}$$

where $G_\ell^{(d)}(k)$ represents the direct path gain for the ℓ^{th} source, $i = \sqrt{-1}$, $\widehat{\boldsymbol{y}}_\ell$ is a unit vector towards the direction of the ℓ^{th} source, and $G_\ell^{(r)}(k, \widehat{\boldsymbol{y}})$ is the reflection gain at the origin along the direction of $\widehat{\boldsymbol{y}}$ for the ℓ^{th} source. Hence, we obtain the spatial domain equivalent of (4.3) by substituting the spatial domain RTF from

(4.4) and (4.5) as

$$P(\boldsymbol{x}_q, k) = \sum_{\ell=1}^{L} S_\ell(k) \left(G_\ell^{(d)}(k)\ e^{ik\ \widehat{\boldsymbol{y}}_\ell \cdot \boldsymbol{x}_q} + \int_{\widehat{\boldsymbol{y}}} G_\ell^{(r)}(k, \widehat{\boldsymbol{y}})\ e^{ik\ \widehat{\boldsymbol{y}} \cdot \boldsymbol{x}_q}\ d\widehat{\boldsymbol{y}} \right) + Z(\boldsymbol{x}_q, k). \tag{4.6}$$

4.3.2 Spherical harmonic decomposition

In this section, we formulate the spherical harmonic expansion of (4.6) based on the harmonic decomposition theory discussed in Section 2.5. Although the subsequent theory is developed for a spherical microphone array, the proposed method is equally effective with various shapes and designs of sensor arrays as long as they meet a few basic criteria of harmonic decomposition [63], [65]–[68], [71].

A continuous function $F(\widehat{\boldsymbol{x}})$ over a sphere can be expressed in the spherical harmonic domain as

$$F(\widehat{\boldsymbol{x}}) = \sum_{nm}^{\infty}\ a_{nm} Y_{nm}(\widehat{\boldsymbol{x}}) \tag{4.7}$$

where $\widehat{\boldsymbol{x}} \equiv (1, \theta, \phi)$ is defined over a sphere, $\sum_{nm}^{(\cdot)} \equiv \sum_{n=0}^{(\cdot)} \sum_{m=-n}^{n}$, $Y_{nm}(\cdot)$ denotes the spherical harmonic of order n and degree m, and a_{nm} indicates corresponding coefficient. Accordingly, the spherical harmonic decomposition of the 3D incident soundfield of (4.6) is given by [1]

$$P(\boldsymbol{x}_q, k) = \sum_{nm}^{\infty}\ \underbrace{\alpha_{nm}(k)\ j_n(kr)}_{a_{nm}(kr)} Y_{nm}(\widehat{\boldsymbol{x}}_q) \tag{4.8}$$

where r is the array radius, $\widehat{\boldsymbol{x}}_q = \boldsymbol{x}_q / r$ is a unit vector towards the direction of the q^{th} microphone, and $\alpha_{nm}(k)$ is the array-independent soundfield coefficient. Eq. (4.8) can be truncated at the soundfield order $N = \lceil ker/2 \rceil$ due to the high-pass nature of the higher order Bessel functions [60], [166], where $e \approx 2.7183$ and $\lceil \cdot \rceil$ denoting the ceiling operation. The soundfield coefficients $\alpha_{nm}(k)$ can be estimated using the technique outlined in Section 2.5.5 where a lower bound $Q \geq (N+1)^2$ needs to be imposed in order to avoid spatial aliasing.

Similarly, the spherical harmonic decomposition of the coherent noise compo-

nent $Z(\boldsymbol{x}_q, k)$ of (4.6) is

$$Z(\boldsymbol{x}_q, k) = \sum_{nm}^{\infty} \eta_{nm}(k) j_n(kr) Y_{nm}(\widehat{\boldsymbol{x}}_q) \tag{4.9}$$

where $\eta_{nm}(k)$ is the soundfield coefficient due to the coherent noise sources. Finally, the spherical harmonic expansion of the Green's function is given by [185, pp. 27–33]

$$e^{ik\,\widehat{\boldsymbol{y}}_\ell \cdot \boldsymbol{x}_q} = \sum_{nm}^{\infty} 4\pi i^n\; Y_{nm}^*(\widehat{\boldsymbol{y}}_\ell)\; j_n(kr)\; Y_{nm}(\widehat{\boldsymbol{x}}_q) \tag{4.10}$$

where $(\cdot)^*$ denotes the complex conjugate operation. Using (4.8), (4.9) and (4.10) in (4.6), we obtain the harmonic-domain representation of a noisy reverberant soundfield by

$$\begin{aligned}\sum_{nm}^{\infty} \alpha_{nm}(k) j_n(kr) Y_{nm}(\widehat{\boldsymbol{x}}_q) = \\ \sum_{nm}^{\infty} \Bigg[4\pi i^n \sum_{\ell=1}^{L} S_\ell(k) \Bigg(G_\ell^{(d)}(k)\; Y_{nm}^*(\widehat{\boldsymbol{y}}_\ell) + \int_{\widehat{\boldsymbol{y}}} G_\ell^{(r)}(k, \widehat{\boldsymbol{y}})\; Y_{nm}^*(\widehat{\boldsymbol{y}})\; d\widehat{\boldsymbol{y}} \Bigg) \\ + \eta_{nm}(k) \Bigg] j_n(kr) Y_{nm}(\widehat{\boldsymbol{x}}_q).\end{aligned} \tag{4.11}$$

Hence, the expression for the combined soundfield coefficients is obtained from (4.11) as

$$\alpha_{nm}(k) = 4\pi i^n \sum_{\ell=1}^{L} S_\ell(k) \Bigg(G_\ell^{(d)}(k)\; Y_{nm}^*(\widehat{\boldsymbol{y}}_\ell) + \int_{\widehat{\boldsymbol{y}}} G_\ell^{(r)}(k, \widehat{\boldsymbol{y}})\; Y_{nm}^*(\widehat{\boldsymbol{y}})\; d\widehat{\boldsymbol{y}} \Bigg) + \eta_{nm}(k) \tag{4.12}$$

$$= \lambda_{nm}(k) + \eta_{nm}(k) \tag{4.13}$$

where $\lambda_{nm}(k)$ is defined as the soundfield coefficients related to the direct and reverberant components of the sound signals.

It is important to note that we consider a far-field sound propagation model in (4.4) and (4.5). For a near-field sound propagation, the corresponding Green's

function and its spherical harmonic expansion is defined as [185, pp. 31]

$$\frac{e^{ik\|\boldsymbol{x}_q-\boldsymbol{y}_\ell\|}}{4\pi\,\|\boldsymbol{x}_q-\boldsymbol{y}_\ell\|}=\sum_{nm}^{\infty}\;ik\;h_n(kr_\ell)\;Y^*_{nm}(\widehat{\boldsymbol{y}}_\ell)\;j_n(kr)\;Y_{nm}(\widehat{\boldsymbol{x}}_q) \tag{4.14}$$

where $\boldsymbol{y}_\ell=(r_\ell,\widehat{\boldsymbol{y}}_\ell)$ is the position vector of ℓ^{th} source and $\|\cdot\|$ denotes the Euclidean distance. In this work, we use the far-field assumption for mathematical tractability, however, the model is equally applicable for a near-field sound propagation.

4.3.3 Spatial coherence of the soundfield coefficients

In this section, we propose novel techniques to develop closed form expressions of the spatial coherence between the harmonic coefficients of reverberant and noise fields in a multi-source environment. From (4.12), the spatial coherence between $\alpha_{nm}(k)$ and $\alpha_{n'm'}(k)$ is

$$\mathbb{E}\Big\{\alpha_{nm}(k)\alpha^*_{n'm'}(k)\Big\}=\mathbb{E}\Big\{\lambda_{nm}(k)\lambda^*_{n'm'}(k)\Big\}+\mathbb{E}\Big\{\eta_{nm}(k)\eta^*_{n'm'}(k)\Big\} \tag{4.15}$$

where we assume uncorrelated speech and noise sources, i.e.

$$\mathbb{E}\Big\{\lambda_{nm}(k)\;\eta^*_{n'm'}(k)\Big\}=0. \tag{4.16}$$

Spatial coherence of the direct and reverberant components

From (4.12) and (4.13), the spatial cross-correlation between the direct and reverberant path coefficients is

$$\begin{aligned}\mathbb{E}\Big\{\lambda_{nm}(k)\lambda^*_{n'm'}(k)\Big\}=C_{nn'}\sum_{\ell=1}^{L}\sum_{\ell'=1}^{L}\;&\mathbb{E}\{S_\ell(k)\;S^*_{\ell'}(k)\}\times\\ &\mathbb{E}\Bigg\{\left(G^{(d)}_\ell(k)\;Y^*_{nm}(\widehat{\boldsymbol{y}}_\ell)+\int_{\widehat{\boldsymbol{y}}}G^{(r)}_\ell(k,\widehat{\boldsymbol{y}})\;Y^*_{nm}(\widehat{\boldsymbol{y}})\;d\widehat{\boldsymbol{y}}\right)\times\\ &\left(G^{(d)*}_{\ell'}(k)\;Y_{n'm'}(\widehat{\boldsymbol{y}}_{\ell'})+\int_{\widehat{\boldsymbol{y}}'}G^{(r)*}_{\ell'}(k,\widehat{\boldsymbol{y}}')\;Y_{n'm'}(\widehat{\boldsymbol{y}}')\;d\widehat{\boldsymbol{y}}'\right)\Bigg\}\end{aligned} \tag{4.17}$$

where $C_{nn'} = 16\pi^2 i^{n-n'}$. Due to the autonomous behaviour of the reflective surfaces in a room (i.e., the reflection gains from the reflective surfaces are independent from the direct path gain), the cross-correlation between the direct and reverberant gains is negligible, i.e.,

$$\mathbb{E}\Big\{ G_\ell^{(d)}(k)\ G_\ell^{(r)*}(k, \widehat{\boldsymbol{y}}) \Big\} = 0. \tag{4.18}$$

Furthermore, we assume that the sources are uncorrelated with each other, and so are the reverberant path gains from different directions, i.e.

$$\mathbb{E}\Big\{ S_\ell(k)\ S_{\ell'}^*(k) \Big\} = \mathbb{E}\Big\{ |S_\ell(k)|^2 \Big\}\ \delta_{\ell\ell'} \tag{4.19}$$

$$\mathbb{E}\Big\{ G_\ell^{(r)}(k, \widehat{\boldsymbol{y}})\ G_\ell^{(r)*}(k, \widehat{\boldsymbol{y}}') \Big\} = \mathbb{E}\Big\{ |G_\ell^{(r)}(k, \widehat{\boldsymbol{y}})|^2 \Big\}\ \delta_{\widehat{\boldsymbol{y}}\widehat{\boldsymbol{y}}'} \tag{4.20}$$

where $|\cdot|$ denotes absolute value. Using (4.18), we eliminate the cross terms of the right hand side of (4.17) and deduce

$$\begin{aligned}\mathbb{E}\Big\{ \lambda_{nm}(k)\lambda_{n'm'}^*(k) \Big\} = C_{nn'} \sum_{\ell=1}^{L}\sum_{\ell'=1}^{L} \mathbb{E}\Big\{ S_\ell(k)\ S_{\ell'}^*(k) \Big\} \\ \Bigg(\mathbb{E}\Big\{ G_\ell^{(d)}(k)\ G_{\ell'}^{(d)*}(k) \Big\}\ Y_{nm}^*(\widehat{\boldsymbol{y}}_\ell)\ Y_{n'm'}(\widehat{\boldsymbol{y}}_{\ell'}) + \\ \int_{\widehat{\boldsymbol{y}}}\int_{\widehat{\boldsymbol{y}}'} \mathbb{E}\Big\{ G_\ell^{(r)}(k, \widehat{\boldsymbol{y}})\ G_{\ell'}^{(r)*}(k, \widehat{\boldsymbol{y}}') \Big\}\ Y_{nm}^*(\widehat{\boldsymbol{y}})\ Y_{n'm'}(\widehat{\boldsymbol{y}}')\ d\widehat{\boldsymbol{y}}\ d\widehat{\boldsymbol{y}}' \Bigg).\end{aligned} \tag{4.21}$$

Defining $\Phi_\ell(k) = \Big(\mathbb{E}\Big\{ |S_\ell(k)|^2 \Big\}\ E\Big\{ |G_\ell^{(d)}(k)|^2 \Big\} \Big)$ as the PSD of the ℓ^{th} source at the origin, we use (4.19) and (4.20) in (4.21) to obtain

$$\begin{aligned}\mathbb{E}\Big\{ \lambda_{nm}(k)\lambda_{n'm'}^*(k) \Big\} = C_{nn'} \sum_{\ell=1}^{L} \Bigg(\Phi_\ell(k)\ Y_{nm}^*(\widehat{\boldsymbol{y}}_\ell)\ Y_{n'm'}(\widehat{\boldsymbol{y}}_\ell) \\ + \mathbb{E}\Big\{ |S_\ell(k)|^2 \Big\} \int_{\widehat{\boldsymbol{y}}} E\Big\{ |G_\ell^{(r)}(k, \widehat{\boldsymbol{y}})|^2 \Big\}\ Y_{nm}^*(\widehat{\boldsymbol{y}})\ Y_{n'm'}(\widehat{\boldsymbol{y}})\ d\widehat{\boldsymbol{y}} \Bigg).\end{aligned} \tag{4.22}$$

Since $|G_\ell^{(r)}(k,\widehat{\boldsymbol{y}})|^2$ is defined over a sphere, we can represent it using the spherical harmonic decomposition as

$$\mathbb{E}\Big\{|G_\ell^{(r)}(k,\widehat{\boldsymbol{y}})|^2\Big\} = \sum_{vu}^{V} \mathbb{E}\Big\{\gamma_{vu}^{(\ell)}(k)\Big\}\, Y_{vu}(\widehat{\boldsymbol{y}}) \tag{4.23}$$

where $\gamma_{vu}^{(\ell)}(k)$ is the coefficient of the power of a reverberant soundfield due to ℓ^{th} source and V is a non-negative integer defining corresponding order. Substituting the value of $\mathbb{E}\{|G_\ell^{(r)}(k,\widehat{\boldsymbol{y}})|^2\}$ from (4.23) into (4.22), we derive

$$\begin{aligned}\mathbb{E}\Big\{\lambda_{nm}(k)\lambda_{n'm'}^*(k)\Big\} = C_{nn'} \sum_{\ell=1}^{L}\Bigg(&\Phi_\ell(k)\; Y_{nm}^*(\widehat{\boldsymbol{y}}_\ell)\; Y_{n'm'}(\widehat{\boldsymbol{y}}_\ell) \\ &+ \mathbb{E}\Big\{|S_\ell(k)|^2\Big\} \sum_{vu}^{V} \mathbb{E}\Big\{\gamma_{vu}^{(\ell)}(k)\Big\} \int_{\widehat{\boldsymbol{y}}} Y_{vu}(\widehat{\boldsymbol{y}})\; Y_{nm}^*(\widehat{\boldsymbol{y}})\; Y_{n'm'}(\widehat{\boldsymbol{y}})\; d\widehat{\boldsymbol{y}}\Bigg).\end{aligned} \tag{4.24}$$

Using the definition of Wigner constants $W_{v,n,n'}^{u,m,m'}$ from Appendix A.1, we rewrite (4.24) as

$$\begin{aligned}&\mathbb{E}\Big\{\lambda_{nm}(k)\lambda_{n'm'}^*(k)\Big\} = \\ &\qquad\sum_{\ell=1}^{L}\Phi_\ell(k)\; C_{nn'}\; Y_{nm}^*(\widehat{\boldsymbol{y}}_\ell)\; Y_{n'm'}(\widehat{\boldsymbol{y}}_\ell) + \sum_{vu}^{V}\Gamma_{vu}(k)\; C_{nn'}\; W_{v,n,n'}^{u,m,m'}\end{aligned} \tag{4.25}$$

where

$$\Gamma_{vu}(k) = \left(\sum_{\ell=1}^{L}\mathbb{E}\{|S_\ell(k)|^2\}\; E\Big\{\gamma_{vu}^{(\ell)}(k)\Big\}\right) \tag{4.26}$$

is the total reverberant power for order v and degree u. Please note that the spatial correlation model developed in [116] was derived for a single source case, i.e. $L = 1$, and did not include background noise in the model.

Spatial correlation model for coherent noise

In a similar way (2.40) was derived, we obtain the expression for η_{nm} from (4.9) as

$$\eta_{nm}(k) = \frac{1}{b_n(kr)} \int_{\widehat{\boldsymbol{x}}} Z(\boldsymbol{x}, k)\ Y_{nm}^*(\widehat{\boldsymbol{x}})\ d\widehat{\boldsymbol{x}} \tag{4.27}$$

where $\boldsymbol{x} = (r, \widehat{\boldsymbol{x}})$. Hence, we deduce

$$\mathbb{E}\Big\{\eta_{nm}(k)\eta_{n'm'}^*(k)\Big\} = \frac{1}{|b_n(kr)|^2} \int_{\widehat{\boldsymbol{x}}} \int_{\widehat{\boldsymbol{x}}'} \mathbb{E}\Big\{Z(\boldsymbol{x}, k)Z^*(\boldsymbol{x}', k)\Big\} Y_{nm}^*(\widehat{\boldsymbol{x}}) Y_{n'm'}(\widehat{\boldsymbol{x}}')\ d\widehat{\boldsymbol{x}}\ d\widehat{\boldsymbol{x}}'. \tag{4.28}$$

The spatial correlation of the coherent noise is given by [197]

$$\mathbb{E}\Big\{Z(\boldsymbol{x}, k)Z^*(\boldsymbol{x}', k)\Big\} = \mathbb{E}\Big\{|Z(\boldsymbol{x}, k)|^2\Big\} \sum_{nm}^{N} j_n(k\,\|\boldsymbol{x} - \boldsymbol{x}'\|)\ 4\pi i^n\ Y_{nm}\left(\frac{\boldsymbol{x}' - \boldsymbol{x}}{\|\boldsymbol{x} - \boldsymbol{x}'\|}\right) \int_{\widehat{\boldsymbol{y}}} \frac{E\left\{|A(\widehat{\boldsymbol{y}})|^2\right\}}{\int_{\widehat{\boldsymbol{y}}} |A(\widehat{\boldsymbol{y}})|^2 d\widehat{\boldsymbol{y}}}\ Y_{nm}^*(\widehat{\boldsymbol{y}})\ d\widehat{\boldsymbol{y}} \tag{4.29}$$

where $A(\widehat{\boldsymbol{y}})$ is the complex gain of the noise sources from $\widehat{\boldsymbol{y}}$ direction. In a reverberant room, the noise field can be assumed to be diffused [198], hence (4.29) reduces to

$$\mathbb{E}\Big\{Z(\boldsymbol{x}, k)Z^*(\boldsymbol{x}', k)\Big\} = \Phi_{z_x}(k)\ j_0(k\,\|\boldsymbol{x} - \boldsymbol{x}'\|) \tag{4.30}$$

where $\Phi_{z_x}(k)$ is the PSD of the noise field at $\boldsymbol{x}$. Furthermore, for the sake of simplicity, we assume that the noise field is spatially white within the small area of a spherical microphone array (e.g., a commercially available spherical microphone array *Eigenmike* [199] has a radius of 4.2 cm), i.e. $\Phi_{z_x}(k) = \Phi_z(k)\ \forall \boldsymbol{x}$. Hence, from (4.28) and (4.30), we get

$$\mathbb{E}\Big\{\eta_{nm}(k)\eta_{n'm'}^*(k)\Big\} = \Phi_z(k) \frac{1}{|b_n(kr)|^2} \int_{\widehat{\boldsymbol{x}}} \int_{\widehat{\boldsymbol{x}}'} j_0(k\,\|\boldsymbol{x} - \boldsymbol{x}'\|)\ Y_{nm}^*(\widehat{\boldsymbol{x}})\ Y_{n'm'}(\widehat{\boldsymbol{x}}')\ d\widehat{\boldsymbol{x}}\ d\widehat{\boldsymbol{x}}'. \tag{4.31}$$

The combined model

Finally, from (4.25) and (4.31), we obtain the complete model of the spatial correlation in a noisy reverberant environment as

$$\mathbb{E}\Big\{\alpha_{nm}(k)\alpha^{*}_{n'm'}(k)\Big\} = \sum_{\ell=1}^{L} \Phi_{\ell}(k)\ \Upsilon^{n'm'}_{nm}(\widehat{\boldsymbol{y}}_{\ell}) + \sum_{vu}^{V}\ \Gamma_{vu}(k)\ \Psi^{m,m',u}_{n,n',v} + \Phi_{z}(k)\ \Omega^{n'm'}_{nm}(k) \tag{4.32}$$

where

$$\Upsilon^{n'm'}_{nm}(\widehat{\boldsymbol{y}}_{\ell}) = C_{nn'}\ Y^{*}_{nm}(\widehat{\boldsymbol{y}}_{\ell})\ Y_{n'm'}(\widehat{\boldsymbol{y}}_{\ell}) \tag{4.33}$$

$$\Psi^{m,m',u}_{n,n',v} = C_{nn'}\ W^{u,m,m'}_{v,n,n'} \tag{4.34}$$

$$\Omega^{n'm'}_{nm}(k) = \frac{1}{|b_n(kr)|^2}\int_{\widehat{\boldsymbol{x}}}\int_{\widehat{\boldsymbol{x}}'} j_0(k\,\|\boldsymbol{x}-\boldsymbol{x}'\|)\ Y^{*}_{nm}(\widehat{\boldsymbol{x}})\ Y_{n'm'}(\widehat{\boldsymbol{x}}')\ d\widehat{\boldsymbol{x}}\ d\widehat{\boldsymbol{x}}'. \tag{4.35}$$

The integrals of (4.35) can be evaluated using a numerical computing tool. An approximation of (4.35) can be made through the finite summations as

$$\Omega^{n'm'}_{nm}(k) \approx \frac{1}{|b_n(kr)|^2}\sum_{q=1}^{Q'}\sum_{q'=1}^{Q'} w_q\ w^{*}_{q'}\ j_0(k\,\|\boldsymbol{x}_q-\boldsymbol{x}_{q'}\|)\ Y^{*}_{nm}(\widehat{\boldsymbol{x}}_q)\ Y_{n'm'}(\widehat{\boldsymbol{x}}_{q'}) \tag{4.36}$$

where $\widehat{\boldsymbol{x}}_q$ and w_q are chosen such a way that the orthonormal property of the spherical harmonics holds. Also, a closed-form expression for (4.35) is derived in Appendix A.2 with the help of the addition theorem of the spherical Bessel functions [184] as

$$\Omega^{n'm'}_{nm}(k) = \frac{(4\pi)^{\frac{3}{2}} i^{(n-n')} j_n(kr) j_{n'}(kr) W^{0,-m,-m'}_{0,n,n'}}{|b_n(kr)|^2}. \tag{4.37}$$

The spatial correlation model of (4.32) is developed considering a far-field sound propagation. Following the discussion of Section 4.3.1, it is evident from (4.10) and (4.14) that a near-field source consideration for the direct path signals changes the direct path coefficient $\Upsilon^{n'm'}_{nm}(\widehat{\boldsymbol{y}}_{\ell})$ of (4.32) as

$$\Upsilon^{n'm'}_{nm}(\widehat{\boldsymbol{y}}_{\ell}, k) = k^2 h_n(kr_{\ell}) h^{*}_{n'}(kr_{\ell})\ Y^{*}_{nm}(\widehat{\boldsymbol{y}}_{\ell})\ Y_{n'm'}(\widehat{\boldsymbol{y}}_{\ell}). \tag{4.38}$$

Hence, to design a system with near-field sources, we require the additional knowledge of the source distance r_ℓ.

4.4 PSD Estimation

In this section, we reformulate (4.32) into a matrix form and solve it in the least square sense to estimate the source, reverberant and noise PSDs. We also discuss an implementation issue and offer engineering solutions to the problem.

4.4.1 Source PSDs

Defining

$$\Lambda_{nm}^{n'm'} = \mathbb{E}\Big\{\alpha_{nm}(k)\alpha_{n'm'}^{*}(k)\Big\}, \tag{4.39}$$

we can write (4.32) in a matrix form by considering the cross-correlation between all the available modes as

$$\boldsymbol{\Lambda} = \boldsymbol{T}\ \boldsymbol{\Theta} \tag{4.40}$$

where

$$\boldsymbol{\Lambda} = [\Lambda_{00}^{00} \quad \Lambda_{00}^{1-1} \dots \Lambda_{00}^{NN} \quad \Lambda_{1-1}^{00} \dots \Lambda_{NN}^{NN}]_{1\times(N+1)^4}^{T} \tag{4.41}$$

$$\boldsymbol{T} = \underbrace{\begin{bmatrix} \Upsilon_{00}^{00}(\widehat{\boldsymbol{y}}_1) & \dots & \Upsilon_{00}^{00}(\widehat{\boldsymbol{y}}_L) & \Psi_{0,0,0}^{0,0,0} & \dots & \Psi_{0,0,V}^{0,0,V} & \Omega_{00}^{00} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Upsilon_{NN}^{NN}(\widehat{\boldsymbol{y}}_1) & \dots & \Upsilon_{NN}^{NN}(\widehat{\boldsymbol{y}}_L) & \Psi_{N,N,0}^{N,N,0} & \dots & \Psi_{N,N,V}^{N,N,V} & \Omega_{NN}^{NN} \end{bmatrix}}_{(N+1)^4\times(L+\{V+1\}^2+1)} \tag{4.42}$$

$$\boldsymbol{\Theta} = [\Phi_1 \dots \Phi_L \quad \Gamma_{00} \dots \Gamma_{VV} \quad \Phi_z]_{1\times(L+\{V+1\}^2+1)}^{T} \tag{4.43}$$

where $(\cdot)^T$ denotes transpose operation. Note that, the frequency dependency is omitted in (4.40)-(4.43) to simplify the notation. Henceforth, we estimate the component PSDs in a *least-square* sense:

$$\widehat{\boldsymbol{\Theta}} = \boldsymbol{T}^{\dagger}\ \boldsymbol{\Lambda} \tag{4.44}$$

where † indicates the pseudo-inversion of a matrix. In a practical implementation, a half-wave rectification or similar measure is required on (4.44) to avoid negative PSDs. The terms Φ_ℓ and Φ_z in the vector $\widehat{\boldsymbol{\Theta}}$ of (4.44) represent the estimated source and noise PSDs at the origin, respectively. It is worth noting that, (4.44) can readily be used for estimating source PSDs in a non-reverberant or noiseless environment by respectively discarding the $\Psi_{n,n',v}^{m,m',u}$ and $\Omega_{nm}^{n'm'}$ terms from the translation matrix $\boldsymbol{T}$ in (4.42).

4.4.2 PSD of the reverberant field

The total reverberation PSD at the origin due to all the sound sources is

$$\Phi_r(k) = \sum_{\ell=1}^{L} \mathbb{E}\Big\{ |S_\ell(k)|^2 \Big\} \int_{\widehat{\boldsymbol{y}}} \mathbb{E}\Big\{ |G_r^{(\ell)}(k, \widehat{\boldsymbol{y}})|^2 \Big\} \, d\widehat{\boldsymbol{y}}. \tag{4.45}$$

Using (4.23), the definition of $\Gamma_{vu}(k)$ in (4.26), and the symmetrical property of the spherical harmonics, (4.45) can be written as

$$\begin{aligned} \Phi_r(k) &= \sum_{vu}^{V} \Gamma_{vu}(k) \int_{\widehat{\boldsymbol{y}}} Y_{vu}(\widehat{\boldsymbol{y}}) \, d\widehat{\boldsymbol{y}} \\ &= \sum_{vu}^{V} \Gamma_{vu}(k) \left(\sqrt{4\pi} \, \delta(v)\delta(u) \right) \\ &= \sqrt{4\pi} \, \Gamma_{00}(k) \end{aligned} \tag{4.46}$$

where $\delta(\cdot)$ is the Dirac delta function. PSD estimation process for a single frequency bin is shown in Algorithm 1.

4.4.3 Bessel-zero issue

One of the challenges in calculating the $\boldsymbol{\Lambda}$ vector is the Bessel-zero issue. We define Bessel-zero issue as the case when $|b_n(kr)|$ of (2.41) takes a near-zero value and thus causes noise amplification and induces error in α_{nm} estimation. This situation arises in 3 distinct scenarios:

Algorithm 1: Algorithm to estimate PSD components

Data: $\boldsymbol{x}_q, P(\boldsymbol{x}_q, k)\ \forall q$

1 Find α_{nm} using (2.41). w_q is manufacture defined;
2 Get $\widehat{y}_\ell\ \forall \ell$ using any suitable DOA estimation technique;
3 Calculate $\Upsilon_{nm}^{n'm'}$, $\Psi_{n,n',v}^{m,m',u}$, and $\Omega_{nm}^{n'm'}$ from (4.33), (4.34) and (4.37), respectively;
4 Get the expected value $\Lambda_{nm}^{n'm'}$ using (4.49);
5 Solve (4.44) for $\boldsymbol{\Theta}$ using the definitions from (4.41) - (4.43).

At low frequencies

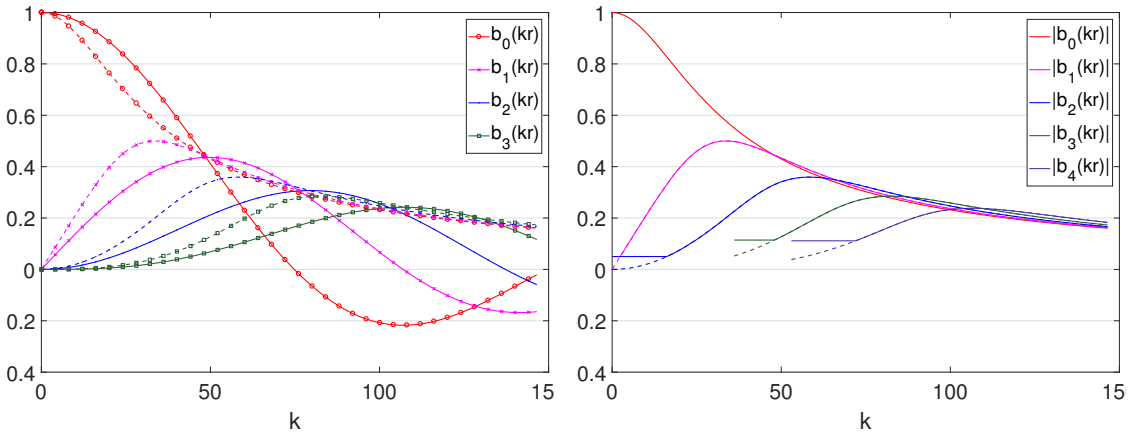
To avoid under-determined solutions as well as to improve the estimation accuracy of (4.44) by incorporating extra spatial modes, we force a minimum value of the soundfield order N at the lower frequency bins. For example, with $V = 3$, $L = 4$ and $f = 500$ Hz, the calculated soundfield order is $N = 1$ and the dimension of $\boldsymbol{T}$ of (4.40) becomes $[16 \times 21]$ which results in an under-determined system. In another scenario, if we choose a smaller value of $V = 1$, though we can avoid an under-determined system, the availability of a fewer spatial modes affects the estimation accuracy. Hence, we impose a lower boundary on N for all frequency bins such that $N = \max\{N, N_{\min}\}$, where $\max\{\cdot\}$ denotes the maximum value and $N_{\min}$ is the lower limit of N. For this work, we choose $N_{\min} = 2$ in an empirical manner. This, however, results in the aforementioned Bessel-zero issue for $n \in [1, N_{\min}]$ at the lower frequencies as shown in Fig. 4.1(a). To avoid this issue, we impose a lower boundary on $b_n(kr)$ as well such that

$$|b_n(kr)| = \max\left\{|b_n(kr)|, b_{n_{\min}}\right\},\ n \in [1, N_{\min}] \tag{4.47}$$

where $b_{n_{\min}}$ is a pre-defined floor value for $|b_n(kr)|$.

At the mode activation boundary

This scenario appears at the first few frequency bins after a higher order mode ($N > N_{\min}$) becomes active. As an example, for $r = 4.2$ cm, 3^{rd} order modes are activated approximately at $k_3^a = 35$ and $k_3^b = 48$, where k_3^a and k_3^b are defined as the values of k when we consider $N = \lceil ker/2 \rceil$ and $N = \lceil kr \rceil$, respectively. In the



(a) Before modification (b) After modification

Figure 4.1: The unaltered Bessel functions with the modified version to alleviate the Bessel-zero issue. (a) Plots unaltered $b_n(kr)$ as a function of k. The complex values are plotted as magnitudes. Solid and dashed lines denote open and rigid arrays, respectively. (b) Shows $|b_n(kr)|$ after modification. Dashed extension denotes the original value.

proposed algorithm, the 3^{rd} order modes are introduced at $k = k_3^a$ and we observe from Fig. 4.1(a) that the value of $|b_3(kr)|$ is close to zero for the first few frequency bins after the activation of the 3^{rd} order modes. To overcome this, we introduce another lower boundary criterion on $|b_n(kr)|$ as

$$|b_n(kr)| = \max\left\{|b_n(kr)|, |b_n(k_n^b r)|\right\},\ n > N_{\min}. \tag{4.48}$$

It is important to note that, the modifications proposed in (4.47) and (4.48) only affect the higher order modes at each frequency bin whereas the lower-order modes remain unchanged. Hence the distortion resulted from these modifications is expected to have less adverse impact than the Bessel-zero issue.

Zero-crossing at a particular frequency

Another case of a Bessel-zero issue occurs when the Bessel functions cross the zero-line on the y-axis at higher frequencies. This is more prominent with the open array configuration as shown in Fig. 4.1(a). The use of a rigid microphone array in the experiment is a way to avoid this issue which we followed in our experiments.

Also note that, the modifications we propose for the previous two scenarios also take care of this zero crossing issue of the Bessel functions for an open array, when $N > 0$.

Fig. 4.1(b) plots the magnitudes of $b_n(kr)$ after the modification for different values of k. The impact of the Bessel-zero issue and the improvement after the proposed modifications are discussed in the result section.

4.5 Experimental Results

In this section, we demonstrate and discuss the experimental results based on practical recordings in a noisy and reverberant room using 4, 6, and 8 speech sources.

4.5.1 Experimental setup

We evaluated the performance in 7 distinct scenarios under 3 different reverberant environments[2] as shown in Table 4.1. The reverberation time T_{60} and DRR in Table 4.1 were calculated based on the methods used in [200]. All the experiments included background noise from the air-conditioning system and the vibration of the electrical equipments in the lab. We created separate training and evaluation datasets that consisted of 320 male and female voices from the TIMIT database [201]. The training dataset was used to set the parameters such as V, $N_{\min}$ etc. whereas the algorithm performance was measured using the evaluation dataset. Each of the 7 scenarios was evaluated 50 times with different mixtures of mixed-gender speech signals making it a total of 350 unique experiments. We used the far-field assumption in each case for computational tractability. We measured the performance with the true and estimated DOA[3] (denoted as "Proposed-GT" and "Proposed-EST", respectively), where the latter was found with a MUSIC-based algorithm [120]. We compared the performance with multiple beamformer-based method of [8] (denoted as "MBF") as, to the best of our knowledge, no similar harmonics-based technique has been proposed in the literature. Note that,

[2] 8-speaker case was not tested in room B & C due to logistical issues.

[3] The true and estimated DOAs for $L = 4, 6$ are listed in Appendix A.3.

for the fairness of comparison, we used all 32 microphones of *Eigenmike* for all the competing methods. Furthermore, as the experiments were designed with the practical recordings instead of a simulation-based approach, the robustness of the proposed algorithm against the realistic thermal noise at the microphones was already evaluated through the process.

Table 4.1: Experimental environments. d_{sm} denotes source to microphone distance.

Room	Dimension (m)	T_{60} (ms)	DRR	d_{sm}	# Speakers
A	$[6.5 \times 4.5 \times 2.75]$	230	10.9 dB	1 m	$4, 6, 8$
B	$[6.5 \times 4.5 \times 2.75]$	230	2.5 dB	2 m	$4, 6$
C	$[11 \times 7.5 \times 2.75]$	640	-0.6 dB	2.8 m	$4, 6$

The *Eigenmike* consists of 32 pressure microphones distributed on the surface of a sphere with a radius of 4.2 cm. The mixed sound was recorded at 48 kHz sampling rate, but downsampled to 16 kHz for computational efficiency. The recorded mixed signals were then converted to the frequency domain with a 8 ms Hanning window, 50% frame overlap, and a 128-point fast Fourier transform (FFT). All the subsequent processing were performed in the STFT domain with the truncated soundfield order $N = 4$, $N_{\min} = 2$, and $b_{n_{\min}} = 0.05$, unless mentioned otherwise. The noise PSD was assumed to have significant power up to 1 kHz whereas all other PSD components were estimated for the whole frequency band. The expected value $\Lambda_{nm}^{n'm'}(k)$ of (4.39) was computed using an exponentially weighted moving average as

$$\Lambda_{nm}^{n'm'}(\tau, k) = \beta \Lambda_{nm}^{n'm'}(\tau - 1, k) + (1 - \beta)\alpha_{nm}(\tau, k)\alpha_{n'm'}^{*}(\tau, k) \tag{4.49}$$

where $\beta \in [0, 1]$ is a smoothing factor, we chose $\beta = 0.8$.

4.5.2 Selection of V

V represents the order of the power of a reverberation soundfield. The exact harmonic analysis of the power of a reverberation soundfield is a difficult task and depends on the structure, orientation, and characteristics of the reflective surfaces.

Hence, unlike a free-field sound propagation, a reverberant field cannot be analytically decomposed into linear combination of Bessel functions which enables the truncation of a non-reverberant soundfield order [31, 32]. Theoretically, V extends to infinity, however, we need to consider several limiting factors such as

- Avoid an under-determined system of equations in (4.40) which imposes a limit on V as
$$V \leq \sqrt{(N+1)^2 - L - 1} - 1. \tag{4.50}$$
- Save computational complexity by choosing the minimum required value of V.

It is also important to note that the nature of the reverberation field plays an important role in determining V. As an example, for a perfectly diffused reverberant room with spatially-uniform reverberant power, only 0^{th} order ($V = 0$) mode is enough. On the other hand, a room with strong directional characteristics requires the higher orders to be considered. Hence, V should be tuned separately for each reverberation environment to obtain an acceptable performance. In our experiments, we chose $V = 0, 4, 8$ for room A, B, and C, respectively, based on the performance with the training dataset.

4.5.3 Evaluation metrics

We evaluate the performance through visual comparisons of the true and estimated PSDs of the sound sources. In addition to that, we also introduce an objective performance index to measure the full-band normalised PSD estimation error as

$$\Phi_{\mathrm{err}_{\ell'}} = 10 \log_{10} \left(\frac{1}{F} \sum_{\forall k} \frac{E\{|\Phi_{\ell'}(\tau, k) - \widehat{\Phi}_{\ell'}(\tau, k)|\}}{E\{|\Phi_{\ell'}(\tau, k)|\}} \right) \tag{4.51}$$

where F is the total STFT frequency bands.

4.5.4 Visualisation of Bessel-zero issue through simulation

In this section, we discuss the practical impact of the Bessel-zero issue, described in Section 4.4.3, on PSD estimation. For this section only, we used a simulated

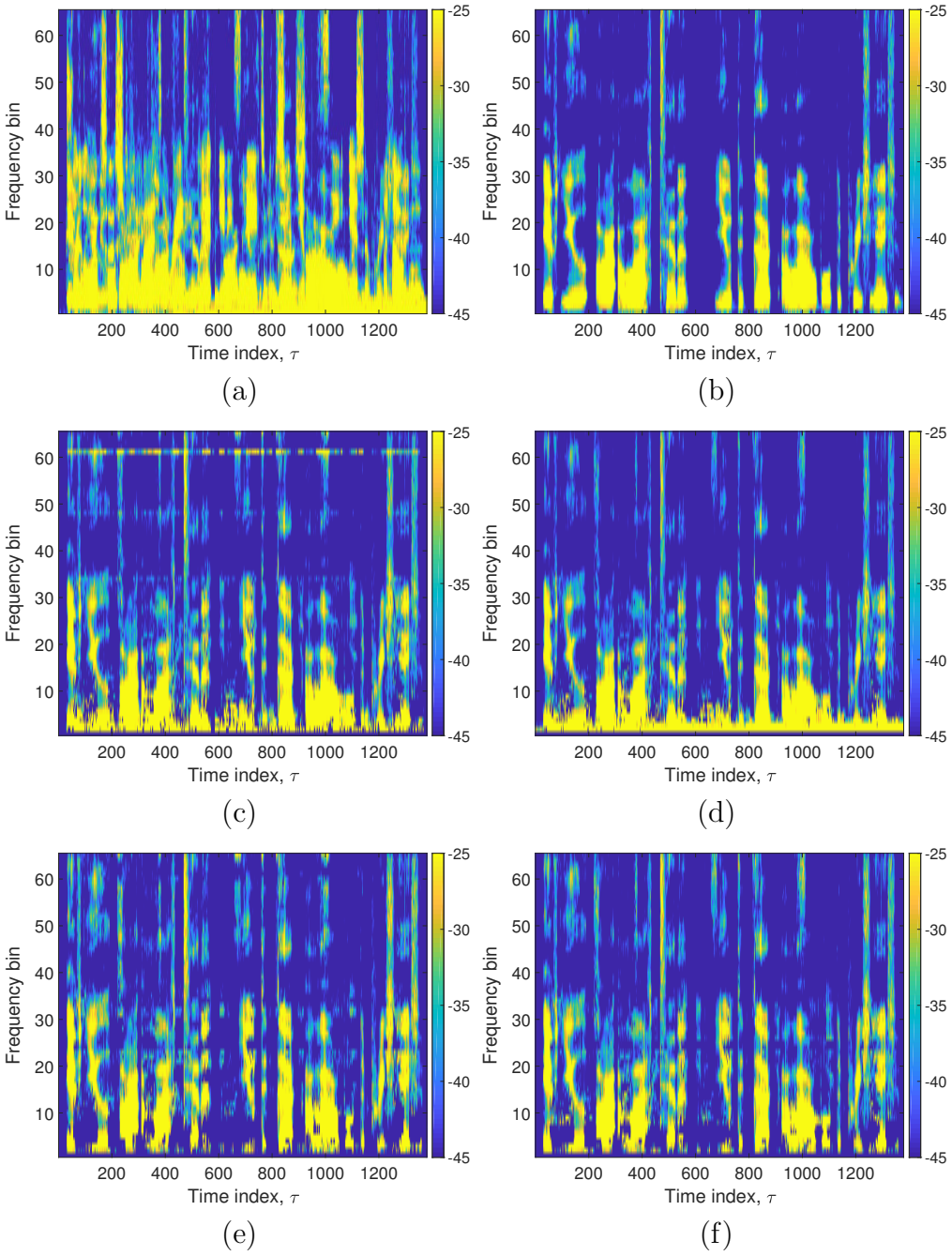
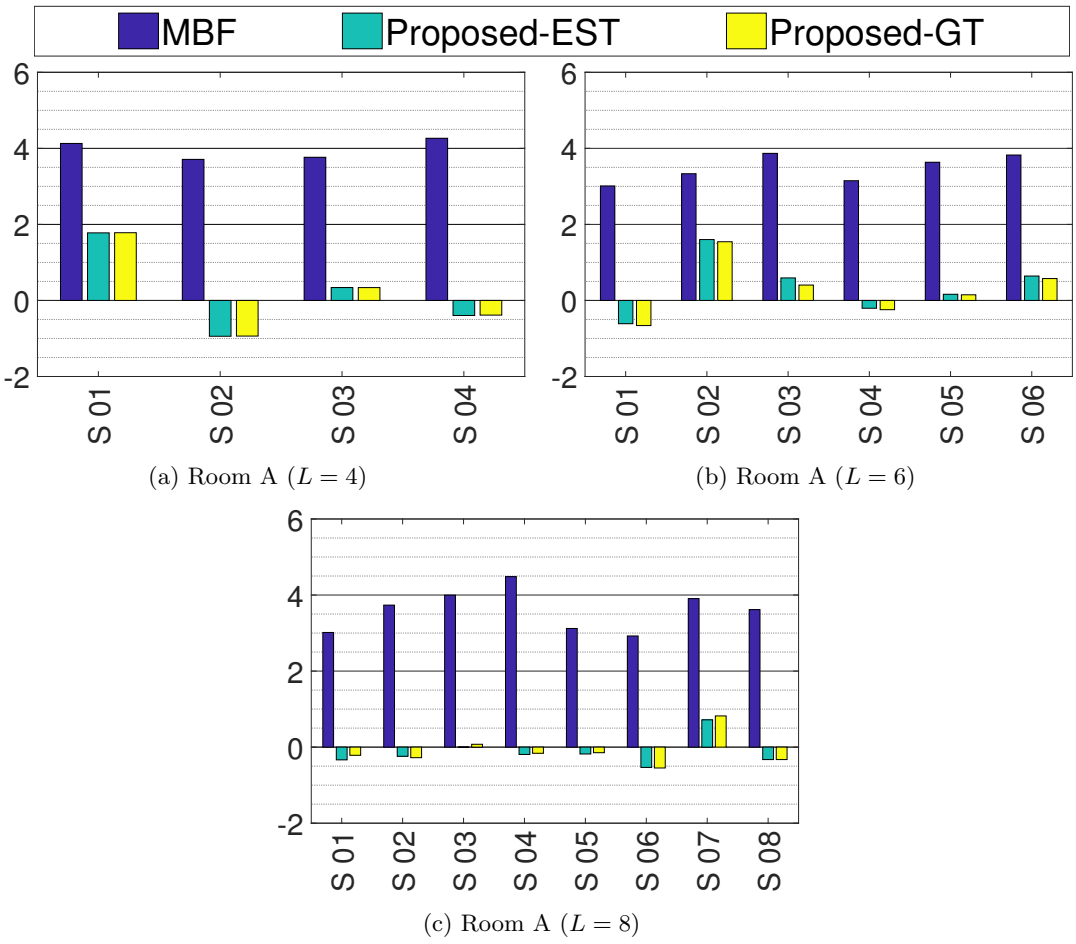


Figure 4.2: The log-spectrograms of the PSDs in a simulated environment to demonstrate the Bessel-zero issue: (a) received signal at the first microphone, (b) true PSD, (c) and (d) estimated PSD without Bessel-zero correction using an open and a rigid array, respectively, and (e) and (f) estimated PSD of with Bessel-zero correction using an open and a rigid array, respectively.

environment to generate the mixed signal, as we required the recordings of both open and rigid arrays to gain a better insight. The simulated environment had identical setup with the practical environment used in the experiments. Fig. 4.2(a) and (b) respectively show the PSDs of the mixed signal and the true PSD of the source signal for S-01. When we estimated the PSD for speaker 1 using an open array and without the proposed Bessel-zero correction (Section 4.4.3), it resulted in Fig. 4.2(c) where the spectral distortion is easily visible at the higher frequencies in the form of isolated horizontal bars (Section 4.4.3) and some random distortions at the lower frequency range (Section 4.4.3). The Bessel-zero issue described in Section 4.4.3 is not prominent here as the impact depends on the spatial location and the relative power of the new incoming mode.

We also tried to solve the Bessel-zero issue by replacing the open array with a rigid array and the result is shown in Fig. 4.2(d). As expected, the rigid array removed the isolated distortions at the higher frequencies, but failed to act on the random distortions at the lower frequency range. It can also be observed that the rigid array resulted an inferior performance in terms of low-frequency noise suppression. As an alternative solution, we used the previous recording from the open array, but this time with the Bessel-zero correction as outlined in Section. 4.4.3. The results, shown in Fig. 4.2(e), provided a better estimation this time by removing most of the Bessel-zero induced distortions. However, few distortions in the form of isolated spectral dots remained at the higher frequencies which were eventually removed when we integrated the proposed solution with a rigid array, as shown in Fig. 4.2(f).

Hence, we conclude that, irrespective of the array type, the most part of the Bessel-zero issue can be overcome through the proposed correction. However, for a better estimation accuracy, it is recommended to integrate the solution with a rigid microphone array. It must be noted that the Bessel-zero corrections can result in some spectral distortion especially at the lower frequency range due to a less number of active modes. However, the gain achieved through these corrections proved to be more significant compared to the resulting distortion, as it will be more evident from the source separation performances we analyse in Chapter 5.



(a) Room A ($L = 4$)

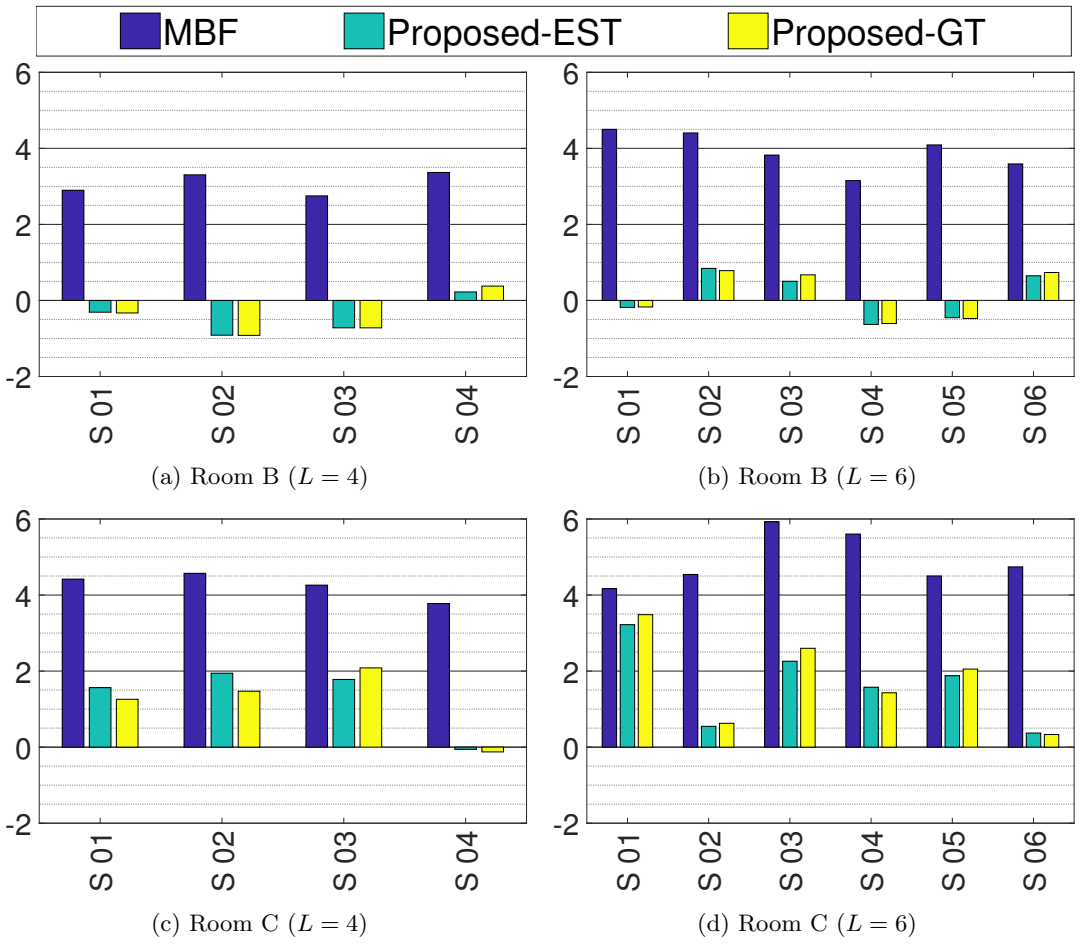
(b) Room A ($L = 6$)

(c) Room A ($L = 8$)

Figure 4.3: Full-band normalised PSD estimation error $\Phi_{\mathrm{err}_{\ell'}}$ in room A (Table 4.1) for different number of sources.

4.5.5 Evaluation of PSD estimation accuracy

Fig. 4.3 and 4.4 show the normalised PSD estimation error in all 7 scenarios where we observe improved PSD estimation for each individual source. In case of room B (Fig. 4.4(a) & (b)) where the source to microphone distance is close to the critical distance, the relative improvement offered by the proposed algorithm is significant which emphasises on the importance of the use of cross-correlation coefficients in highly reverberant environments. We also observe notable improvements in room C (Fig. 4.4(c) & (d)) where we have a weaker direct path compared to the reverberant path (DRR $<$ 0 dB). However, the performance in room C was affected due to



(a) Room B ($L = 4$)

(b) Room B ($L = 6$)

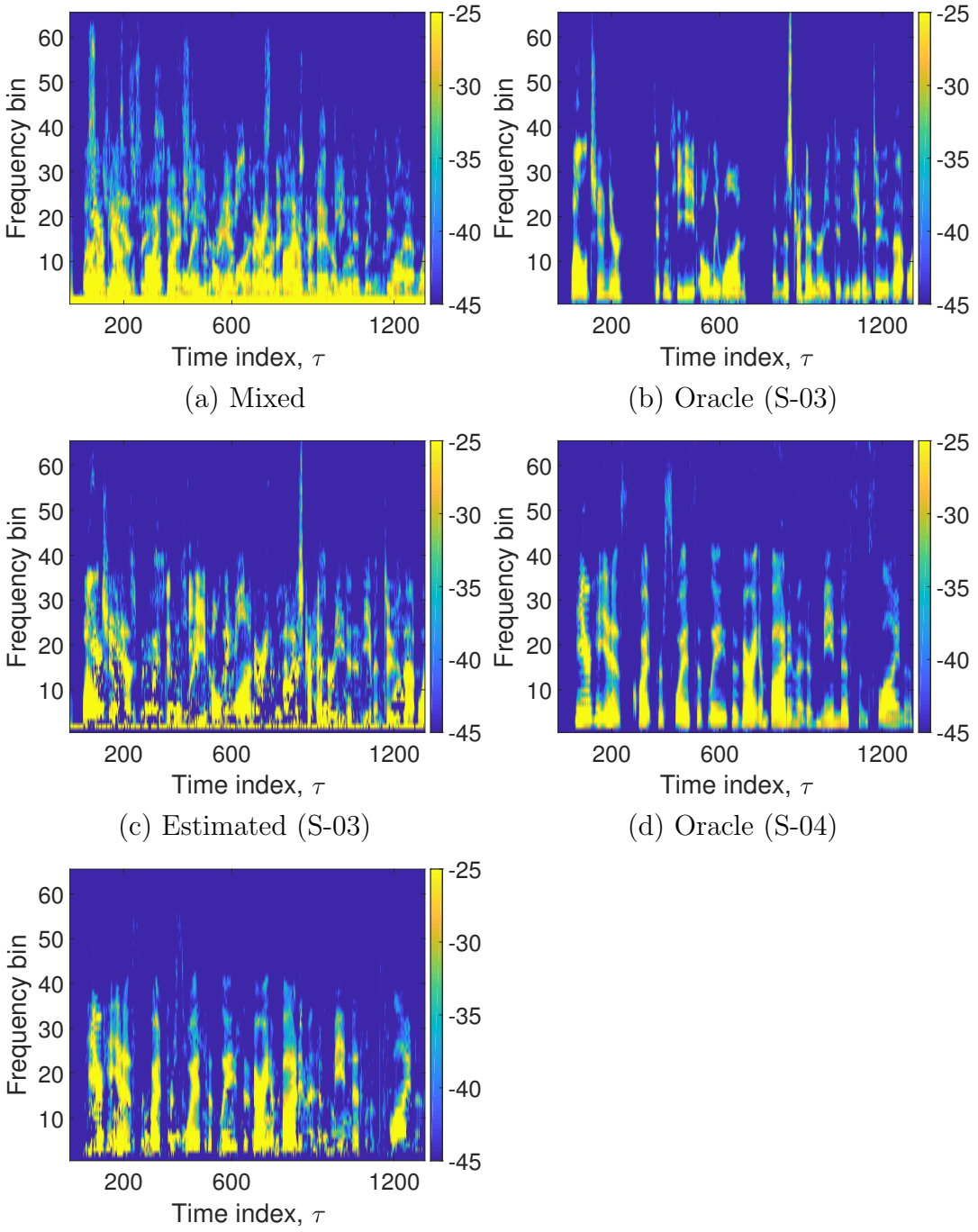
(c) Room C ($L = 4$)

(d) Room C ($L = 6$)

Figure 4.4: Full-band normalised PSD estimation error $\Phi_{\text{err}_{\ell'}}$ in room B and C (Table 4.1) for different number of sources.

the non-uniform reflective surfaces (e.g. glass and brick walls) which resulted in relatively strong directional characteristics. This could be improved if the order V was allowed to be increased which was not possible due to (4.50). Finally, with the DOA estimation accuracy within 4 degree (Table A.1), no major performance deviation was observed for true and estimated DOA consideration.

Fig. 4.5 shows the original and the estimated PSDs in room A for S-03 and S-04 along with the mixed signal PSD for 4-speaker case. From Fig. 4.5 we observe that S-04 estimation exhibits a very good resemblance to the original signal whereas S-03 are affected by few spectral distortions. This is due to the relative difference in signal strength in terms of signal to noise ratio (SNR) and signal to



(a) Mixed

(b) Oracle (S-03)

(c) Estimated (S-03)

(d) Oracle (S-04)

(e) Estimated (S-04)

Figure 4.5: The log-spectrograms of the estimated PSDs for a 4-source setup in room A. The received signal PSD at microphone 1 and the true PSDs are included for reference.

interference ratio (SIR) as S-03 possessed the lowest values of SNR and SIR among all the sources. We also notice the presence of very low frequency background noise in the estimated PSD of S-03. This can be a result of the spatial position of S-03 in addition to the aforementioned SNR and SIR issues. This problem can be resolved by pre-filtering the input signal with a high pass filter (HPF) to remove the signal below 200 Hz. Furthermore, few random spectral distortions are observed in some frequencies which are mainly contributed by the practical limitations such as source and microphone positioning error, Bessel-zero correction, the deviation of the speaker and microphone characteristics from the ideal scenario, the finite correlation between the sources and the reverberation components due to limited STFT window length and imperfect room characteristics, measurement inaccuracies etc.

4.6 Summary

The main objective of this chapter was to develop a mathematical model for the modal coherence of a noisy and reverberant soundfield in order to exploit that for overcoming various signal processing challenges such as PSD estimation, source separation, and DOA estimation. To that end, we achieved the following outcomes from this chapter

- We developed a mathematical model for the modal coherence of a multi-source reverberant soundfield using its spherical harmonic coefficients.
- We derived a novel closed-form expression of coherent noise field by using its spherical harmonic coefficients.
- We utilised the modal coherence model to estimate power spectral densities of individual sound components of a complex noisy and reverberant soundfield. The modal coherence model is capable of extracting a significantly larger number of sources compared to the conventional beamformer-based solutions.
- We analysed and investigated various implementation challenges including the impact of the Bessel-zero issue and offered engineering solutions to them.

- We measured PSD estimation accuracy in various practical environments with a commercial microphone array. The relative comparison revealed that the proposed method outperformed other competing methods under diverse acoustic environments.

Based on the foundation of the modal coherence model and PSD estimator we developed so far, in the next chapter we attempt to design and evaluate a source separation methodology using complete as well as partial coherence matrices.

4.7 Related Publications

- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "PSD Estimation and Source Separation in a Noisy Reverberant Environment using a Spherical Microphone Array", IEEE/ACM Transactions on Audio Speech and Language Processing, Volume 26, Issue 9, pp. 1594–1607, 2018.

- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "PSD Estimation of Multiple Sound Sources in a Reverberant Room Using a Spherical Microphone Array", Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 76–80, New York, USA, October 2017.

This page intentionally left blank.

Chapter 5

Application of Modal Coherence-based PSD Estimation in Source Separation

Source separation is a common requirement in many acoustic signal processing applications. Despite having a long research history, this is still an active research area due to various limitations and constraints of the existing algorithms considering the complex dynamic nature of audio signals. In this chapter, we investigate the application of the modal coherence-based PSD estimation technique we outlined in the previous chapter by performing source separation in a noisy and reverberant environment. We demonstrate two different approaches to the solution. First, we use a spherical microphone array to extract the full coherence matrix and subsequently attain source separation using the estimated PSDs. Next, we propose an extension of this method by designing a planar array which utilises only a partial coherence matrix in order to reduce the computational cost and hardware requirement.

5.1 Introduction

In this chapter, we explore a practical application of the modal coherence model we developed in Chapter 4 in the form of source separation. Source separation is a useful technique in many signal processing applications [19], [202], [203]. Various techniques have been developed in literature to accomplish a reliable source separation technique [76], [83], [91], [99]. However, each technique exhibits certain advantages as well as limitations under specific acoustic scenarios. Hence, various constraints and priors need to be considered for the existing techniques that keep the source separation an active research area. In this chapter, we evaluate source separation performance based on the modal coherence model we developed in the previous chapter and compare it with the contemporary solutions.

First we analyse the performance of the algorithm exploiting the full coherence matrix employing a spherical microphone array. We conduct multiple experiments in several practical room environments with a commercially available microphone array, *Eigenmike* [199], without any prior knowledge about the source characteristics. We validate the performance of the algorithm by carrying out 350 experiments in 3 different acoustic environments with varying number of speakers using mixed-gender speech signals. The performance is evaluated in terms of perceptual evaluation of speech quality (PESQ) [204] and frequency-weighted segmental signal to noise ratio (FWSegSNR) [193] and compared against multiple contemporary methods.

Later we investigate an alternative array structure to achieve spatial coherence-based source separation using a planar microphone array. We only expose a partial coherence matrix to achieve the desired outcome with a significantly smaller number of microphones and a simpler array structure. This is useful in integrating the solution with various commercial products such as smart home appliances. The performance of the planar array-based source separation is compared with different existing techniques as well as spherical and differential microphone arrays.

The rest of the chapter is organised as follows. Section 5.2 defines the problem statement for this chapter as a continuation of the methods we developed in Chapter 4. In Section 5.3, we describe the source separation algorithm based on the full modal coherence matrix. A detailed experimental validation of Section 5.3 is

demonstrated in Section 5.4 using a spherical microphone array. In Section 5.5, the theory modal coherence-based source separation is modified to use with a planar array. Finally, the performance of the planar array with the proposed model is analysed in Section 5.6.

5.2 Problem Statement

To avoid repetition, we start from where we left in the previous chapter. We already established in Section 4.3 and 4.4 that the power spectral density (PSD) of the individual audio components of a noisy and reverberant soundfield can be estimated by solving the following system of equations derived from the modal coherence model:

$$\begin{bmatrix} \Lambda_{00}^{00} \\ \vdots \\ \Lambda_{00}^{NN} \\ \Lambda_{1-1}^{00} \\ \vdots \\ \Lambda_{NN}^{NN} \end{bmatrix} = \underbrace{\begin{bmatrix} \Upsilon_{00}^{00}(\widehat{\boldsymbol{y}}_1) & \dots & \Upsilon_{00}^{00}(\widehat{\boldsymbol{y}}_L) & \Psi_{0,0,0}^{0,0,0} & \dots & \Psi_{0,0,V}^{0,0,V} & \Omega_{00}^{00} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Upsilon_{00}^{NN}(\widehat{\boldsymbol{y}}_1) & \dots & \Upsilon_{00}^{NN}(\widehat{\boldsymbol{y}}_L) & \Psi_{0,N,0}^{0,N,0} & \dots & \Psi_{0,N,V}^{0,N,V} & \Omega_{00}^{NN} \\ \Upsilon_{1-1}^{00}(\widehat{\boldsymbol{y}}_1) & \dots & \Upsilon_{1-1}^{00}(\widehat{\boldsymbol{y}}_L) & \Psi_{1,0,0}^{-1,0,0} & \dots & \Psi_{1,0,V}^{-1,0,V} & \Omega_{1-1}^{00} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Upsilon_{NN}^{NN}(\widehat{\boldsymbol{y}}_1) & \dots & \Upsilon_{NN}^{NN}(\widehat{\boldsymbol{y}}_L) & \Psi_{N,N,0}^{N,N,0} & \dots & \Psi_{N,N,V}^{N,N,V} & \Omega_{NN}^{NN} \end{bmatrix}}_{\boldsymbol{T}} \begin{bmatrix} \Phi_1 \\ \vdots \\ \Phi_L \\ \Gamma_{00} \\ \vdots \\ \Gamma_{VV} \\ \Phi_z \end{bmatrix} \tag{5.1}$$

where $\Lambda_{nm}^{n'm'} = \mathbb{E}\Big\{\alpha_{nm}(k)\alpha_{n'm'}^*(k)\Big\}$ is based on the measurements, Υ, Ψ, and Ω are design parameters given by (4.33), (4.34), and (4.37), respectively. The source PSDs Φ_ℓ $\forall \ell \in [1, L]$ for L active sources and noise PSD Φ_z are readily attainable from (5.1) using a *least-square* method. Finally, PSD of the reverberant field can be calculated from the following equation

$$\Phi_r(k) = \sqrt{4\pi}\ \Gamma_{00}(k). \tag{5.2}$$

Hence, we define our problem for this chapter as to estimate source signal $S_\ell(k)$ $\forall \ell$ given the measured sound pressure $P(\boldsymbol{x}_q, k)$ $\forall q \in [1, Q]$, where Q is the number of microphones, or the corresponding spherical harmonic coefficients $\alpha_{nm}(k)$.

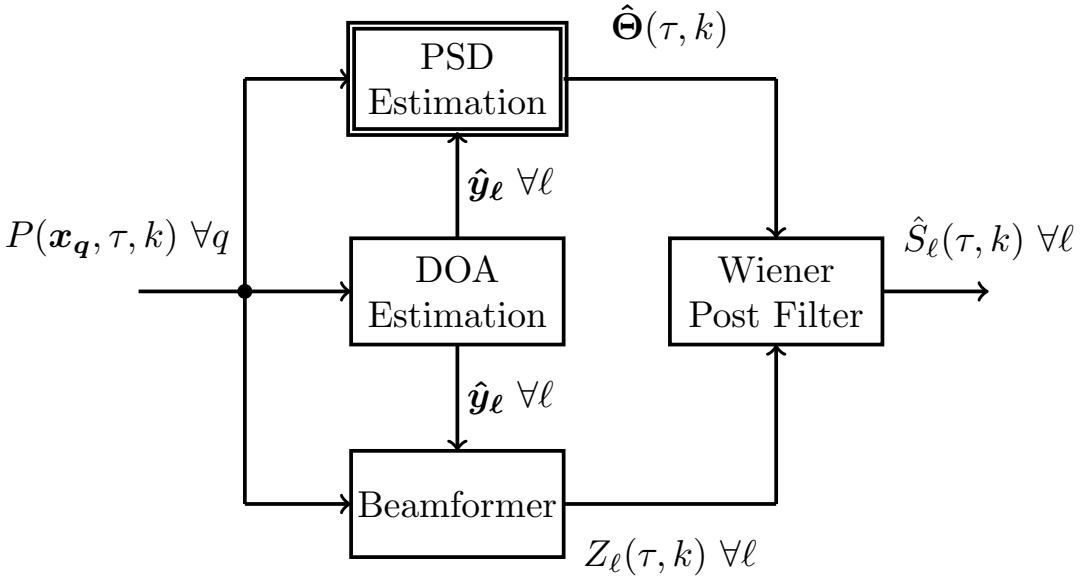


Figure 5.1: Block diagram of an application of the proposed PSD estimation method in terms of source separation.

5.3 Source Separation using Full Modal Coherence Matrix

In this section, we perform PSD estimation based on the full modal coherence matrix utilising all available soundfield modes up to order N. The estimated PSDs then are used to design a Wiener filter to boost the interference rejection of a beamformer output. The performance of the Wiener filter largely depends on the estimation accuracy of the source and interfering PSDs, which is where the importance of a PSD estimation algorithm lies. A block diagram of the complete source separation methodology is shown in Fig. 5.1 and explained in the subsequent sections.

5.3.1 Estimation of the direction of arrival

The PSD estimator as well as the beamformer requires the knowledge of the directions of arrival (DOA) of sound sources. If the source positions are unknown,

a localisation technique, e.g., multiple signal classification, commonly known as MUSIC [12], has to be used to estimate the DOA of the source signals. For the current work, we focus on measuring the performance of source separation, hence, we demonstrate the evaluation results using both oracle DOA as well as single-source DOA estimation based on a frequency-smoothed approach of the MUSIC algorithm [120].

5.3.2 Choice of beamformer

There are several beamforming techniques available in literature such as delay and sum (DS), maximum directivity (MD), or minimum variance distortionless response (MVDR) etc. The choice of the beamforming technique depends on the application and objective of the exercises. In this work where the undesired signal includes the correlated reverberant component of the desired signal, an MVDR beamformer can result desired signal cancellation if the undesired PSD components at each microphone position are unknown. Hence, a simple delay and sum beamformer or a maximum directivity beamformer is more appropriate for the current case whose output, when steered towards ℓ^{th} far-field source, is given by [63], [205]

$$\widehat{S}_{bf}^{(\ell)}(k) = \sum_{nm}^{N} d_n(kr)\ \alpha_{nm}(k) Y_{nm}(\theta_\ell, \phi_\ell) \tag{5.3}$$

where

$$d_n(kr) = \begin{cases} \frac{i^{-n}}{(N+1)^2} & \text{for an MD beamformer} \\ \frac{4\pi |b_n(kr)|^2}{i^n} & \text{for a DS beamformer} \end{cases}. \tag{5.4}$$

5.3.3 Wiener post-filter

Regardless of the choice of beamformer, a post filter is known to enhance the beamformer output in most of the cases [11], [106], [198]. Hence, at the last stage, we apply a Wiener post filter at the beamformer output using the estimated PSDs.

The transfer function of a Wiener filter for the ℓ^{th} source is given by

$$H_w^{(\ell)}(k) = \frac{\Phi_\ell(k)}{\sum_{\ell'=1}^{L} \Phi_{\ell'}(k) + \Phi_r(k) + \Phi_z(k)}. \tag{5.5}$$

where all the PSD components are already estimated by the proposed algorithm and available in the vector $\widehat{\mathbf{\Theta}}$. Hence, the ℓ^{th} source signal is estimated by

$$\widehat{S}_\ell(k) = \widehat{S}_{bf}^{(\ell)}(k)\ H_w^{(\ell)}(k). \tag{5.6}$$

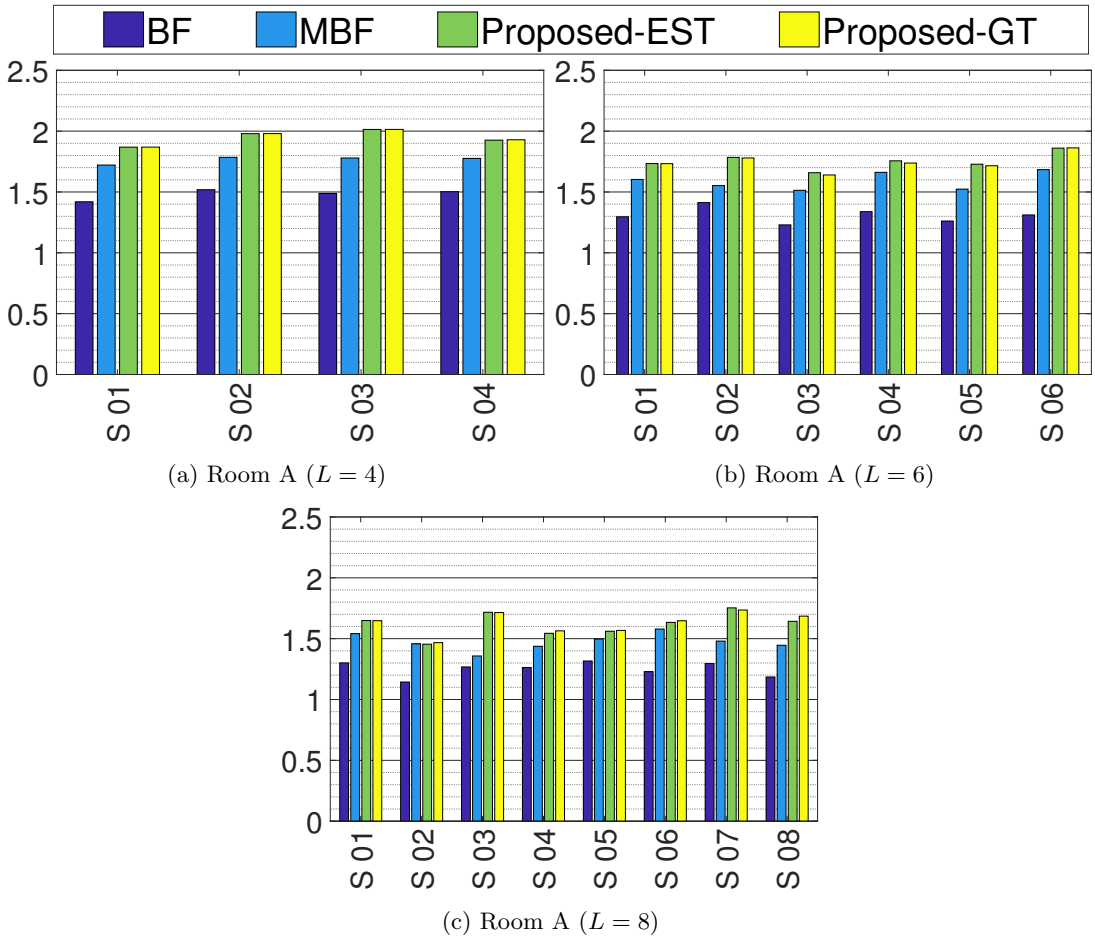
5.4 Experiments using a Spherical Array

Based on the foregoing discussion, we designed several experiments based on practical measurements using a commercial microphone array. In this section, we present the source separation performance in 3 different room conditions of Table 5.1 based on the same dataset we used in Chapter 4. The experimental setup remains identical to the one described in Section 4.5.1. Note that, all the measurements were taken in practical acoustic scenarios, hence, the experiments captured all the physical deviations including background noise, microphone noise, and measurement inaccuracies.

Table 5.1: Experimental environments. d_{sm} denotes source to microphone distance.

Room	Dimension (m)	T_{60} (ms)	DRR	d_{sm}
A	$[6.5 \times 4.5 \times 2.75]$	230	10.9 dB	1 m
B	$[6.5 \times 4.5 \times 2.75]$	230	2.5 dB	2 m
C	$[11 \times 7.5 \times 2.75]$	640	-0.6 dB	2.8 m

For the comparative performance evaluation, we retain the three methods we used in Section 4.5, namely multiple beamformer-based "MBF" [8], oracle DOA knowledge-based "Proposed-GT", and estimated DOA-based "Proposed-EST". In



(a) Room A ($L = 4$)

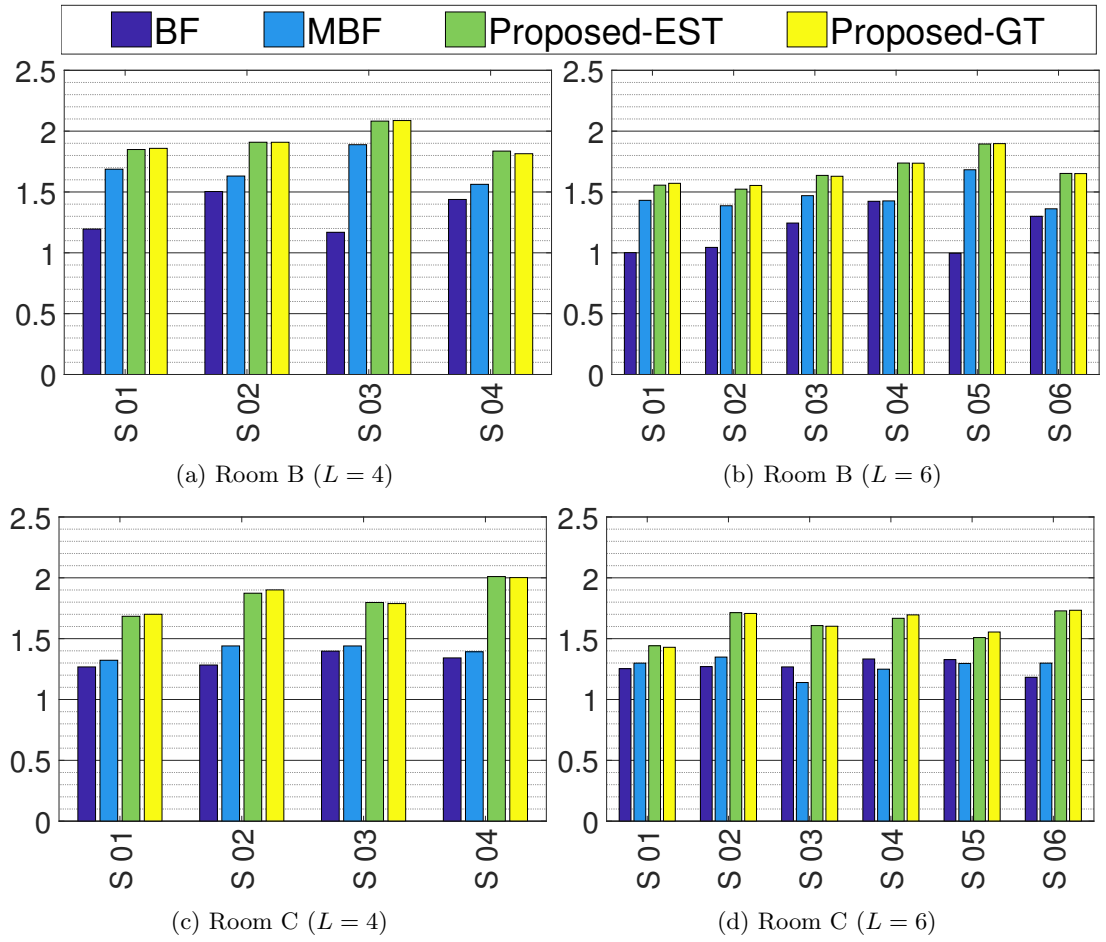
(b) Room A ($L = 6$)

(c) Room A ($L = 8$)

Figure 5.2: PESQ in room A (Table 5.1) for different number of sources.

addition to that, we include the results from the beamformer output (denoted as "BF") in order to demonstrate the improvement offered by the post-filtering block.

The source separation performances are measured frequency-weighted segmental signal to noise ratio (FWSegSNR) [193] and perceptual evaluation of speech quality (PESQ) [204]. Each of the experiment was performed 50 times with random mixture of speech and the average results from all the experiments are presented in the subsequent sections.



(a) Room B ($L = 4$)

(b) Room B ($L = 6$)

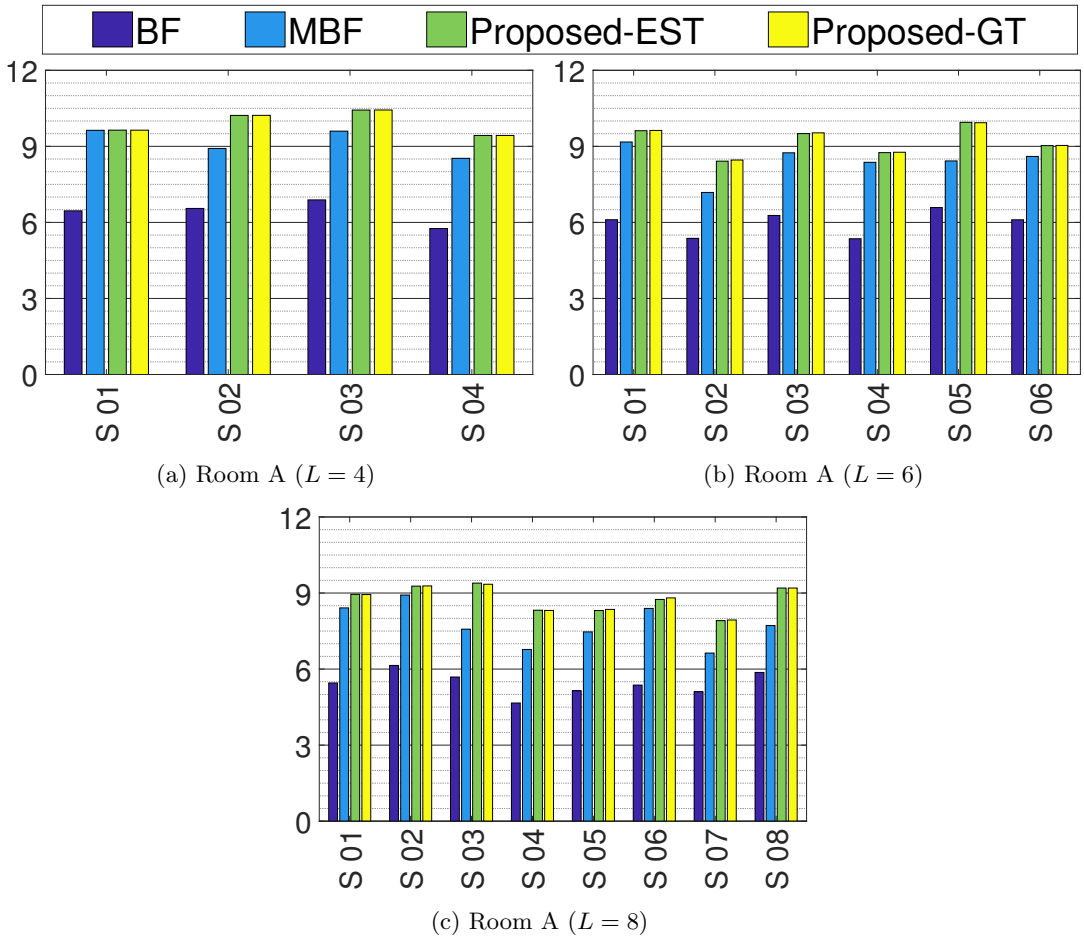
(c) Room C ($L = 4$)

(d) Room C ($L = 6$)

Figure 5.3: PESQ in room B and C (Table 5.1) for different number of sources.

5.4.1 Performance evaluation of source separation

We first investigate the impact based on PESQ, which is an ITU-T recommended metric to measure the overall quality, speech distortion, and noise distortion. It was shown in [193] that PESQ exhibits the highest correlation with overall quality and signal distortion among the widely-used objective metrics. Fig. 5.2 and 5.3 plot PESQ in all 7 acoustic scenarios for the competing methods. It is obvious from the plots that the proposed method outperforms the beamformer output as well as "MBF" by a large margin under each scenario. We also notice that the improvement offered by "MBF" over the traditional beamformer diminishes as the number of sources increases. This is due to the heuristic selection of beamformer directivity



(a) Room A ($L = 4$)

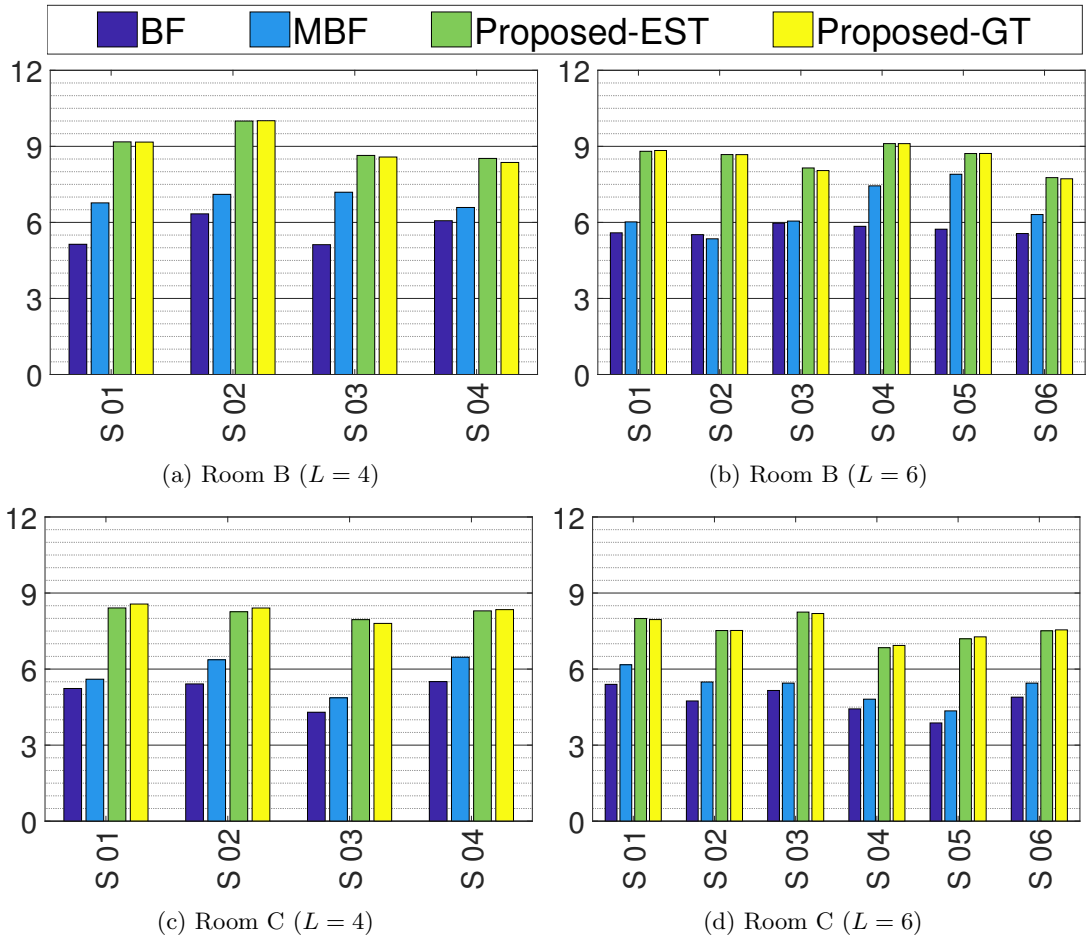
(b) Room A ($L = 6$)

(c) Room A ($L = 8$)

Figure 5.4: FWSegSNR (dB) in room A (Table 5.1) for different number of sources.

in [8] that resulted in ill-posed problem as the number of sources increases, and hence, affects its performance. The performance issue with "MBF" is found to be more prominent under strong reverberation (Fig. 5.3) compared to Fig. 5.2 where direct path is significantly stronger with DRR of 10.9 dB. Conversely, irrespective of oracle or estimated DOA, the proposed algorithm avoids such distortion due to the well-structured nature of orthogonal spherical harmonic basis functions that ensures the maximum spatial dissociation among the harmonic beams.

A similar trend is observed in Fig. 5.4 and 5.5 which plot FWSegSNR for the competing methods. FWSegSNR measures the spectral similarities between the competing methods and correlates well with overall quality and background



(a) Room B ($L = 4$)

(b) Room B ($L = 6$)

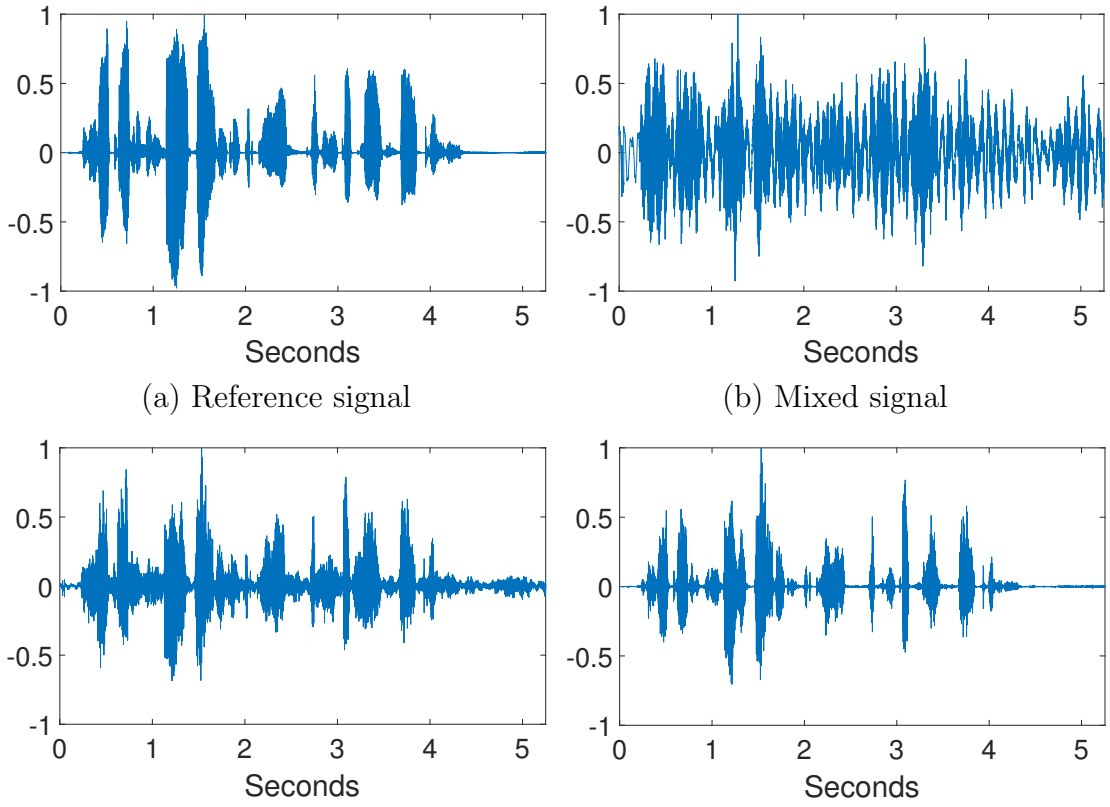
(c) Room C ($L = 4$)

(d) Room C ($L = 6$)

Figure 5.5: FWSegSNR (dB) in room B and C (Table 5.1) for different number of sources.

distortion of an audio signal [193]. Similar to PESQ, FWSegSNR also confirms the superior performance of the proposed algorithm in each room condition. The performance of the proposed method remains comparatively unaffected against stronger reverberation whereas "MBF" suffers from the presence of stronger reflections. This is credited by the fact that the inclusion of the cross-terms in the proposed algorithm significantly increases the spectral resolution of the dissection and achieves better outcomes in rejecting the undesired reflections.

The source separation performance shown here agrees with the PSD estimation accuracy demonstrated in Section 4.5.5. This further establishes the fact that the MSE error shown in Fig. 4.3 and 4.4 of previous chapter were within an acceptable



(a) Reference signal (b) Mixed signal

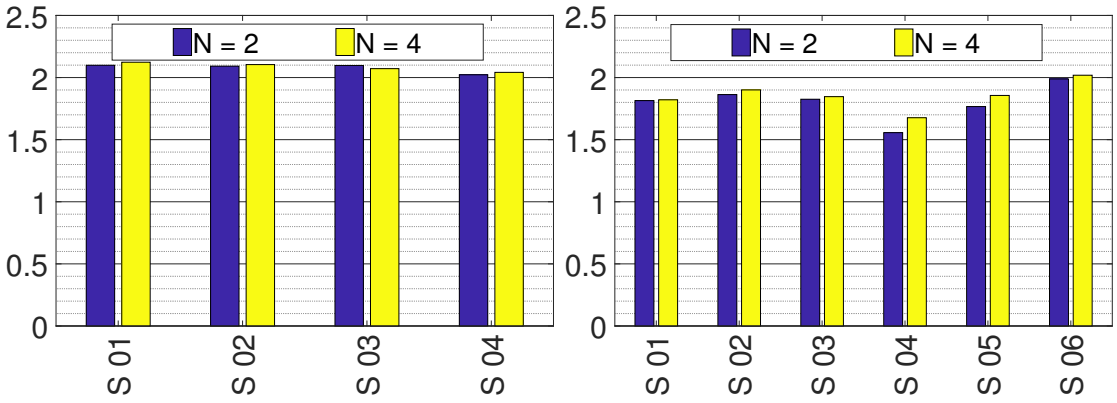
(c) Beamformer output (d) Proposed method

Figure 5.6: The estimated waveforms of speaker 1 in Room A. The waveform at the beamformer output along with the original and the mixed signal waveforms are shown for reference.

range for source separation application. Furthermore, the notable improvement in terms of PESQ confirms minimum signal distortion due to the Bessel zero correction we incorporated during the PSD estimation task.

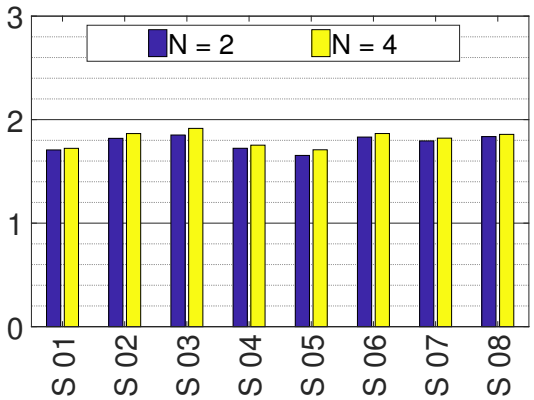
It is worth noting that, our primary objective was to measure the performance of modal coherence-based source separation technique and compare it with the contemporary algorithms. Hence, we did not make any considerable effort in relation to the design of the beamformer block in Fig 5.3.1, instead, we ensured that the same beamformer design was followed throughout the experiments.

Finally in Fig. 5.6, we present time domain acoustic snapshots of a speaker in Room A captured at different stages of Fig. 5.1 and compare it with the reference. It is clear from the plot that, while a beamformer can only partially restore the



(a) 4-speaker case

(b) 6-speaker case



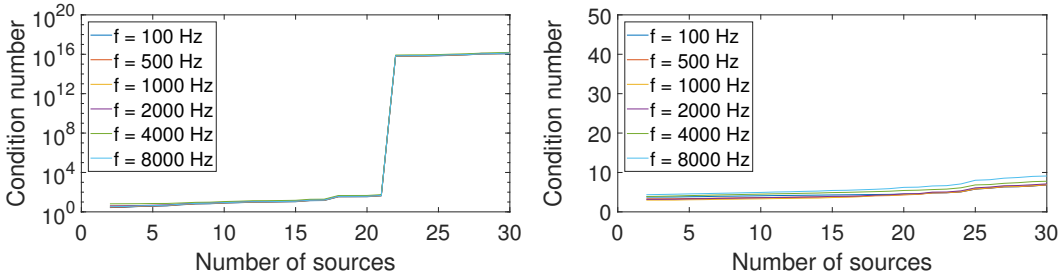
(b) 8-speaker case

Figure 5.7: PESQ in Room A for estimated source signals with $N = 2$ and 4.

original signal, a Wiener post-filter significantly improves the quality given that we achieve an accurate estimation of the signal and interfering PSDs.

5.4.2 Impact of array size and order on system performance and error sensitivity

One of the major challenges in spherical harmonics-based solutions is the number of microphones required to calculate all soundfield coefficients. The required number of microphones is directly related to the maximum soundfield order N. Theoretically, to calculate the spherical harmonic coefficients of an N^{th}-order soundfield, we generally require at least $(N+1)^2$ microphones. So far, we have used $N = 4$ in our



(a) $N = 2$ (b) $N = 4$

Figure 5.8: Condition number of the transfer matrix $\boldsymbol{T}$ with $N = 2$ and 4.

experiments; however, reducing it to $N = 2$ does not have any significant adverse impact on the performance, as shown in Fig. 5.7. Hence, for the demonstrated examples, it is possible to utilise a lower-order microphone array [66] without having a major performance degradation.

However, as the number of sources increases or the room exhibits lesser diffused reverberation, a higher order decomposition ensures a better estimation accuracy as it offers more accurate knowledge of the spatial distribution of soundfield. Furthermore, additional modes help to avoid ill-posed problem of (5.1) in the presence of a larger number of sources. This is evident from Fig. 5.8 which plots the condition number of $\boldsymbol{T}$ for $N = [2, 4]$ with $V = 1$ against different number of sources and frequencies. The sources considered in Fig. 5.8 were uniformly distributed on the surface of a 1 m sphere at 5 different azimuth planes. For the case of $N = 2$, the condition number of $\boldsymbol{T}$ remains low up to 21 sources, but increases rapidly beyond that. On the contrary for $N = 4$, the system remains well-posed within the experimental limit of 30 sources. Notably, the same behaviour is observed over a wide frequency range. This is due to the fact that only the noise terms in the last column of $\boldsymbol{T}$ are frequency dependent when the far-field assumption is made.

5.5 A Planar Array for Source Separation

In practical applications, we often seek to reduce computational complexity and manufacturing cost while exploring seamless integration possibilities, even at the cost of imposing certain constraints. In this section, we explore such an option to

apply the same modal coherence-based source separation that we described earlier, but with a simpler planar array.

5.5.1 Motivation for a planar array

The motivation for a simple planar array comes from the fact that the PSD estimation and source separation technique we discussed so far requires a minimum $(N+1)^2$ microphones when used with a spherical microphone array [63], [71] or $\big(2(N+1)^2-2\big)$ omni-directional microphones with a hybrid differential microphone array [65]. A reduced number of microphones is desirable from many commercial perspectives, especially when a smaller number of sources are to be considered. Furthermore, a planar array offers a reduced design complexity compared to a spherical microphone array and can easily be integrated in different practical devices such as smart home appliances. Hence, we explore the idea of using a planar microphone array in performing the source separation algorithm exploiting certain properties of spherical harmonics.

Based on the fact that the modal coherence model for PSD estimation, shown in (5.1), is array-independent and fully scalable, the sufficient criteria to solve (5.1) is

$$N_{\text{coeff}} \geq L + (V+1)^2 + 1 \tag{5.7}$$

where N_{coeff} is the number of coherence coefficients. Frequently, we end up having an over-determined system of (5.1), hence, it is possible to discard a few modes without significantly impacting the performance of the algorithm. In the subsequent sections, we systematically pick a subset of the available modes to construct (5.1) with a reduced number of microphones.

We also consider eliminating the beamformer of Fig. 5.1 to reduce storage and computational requirements. Furthermore, as we discussed in Section 2.5.3, the far-field approximation of Green's function is normalised with respect to the source signal at the origin. Hence, the definition of the PSD terms Φ_ℓ and Γ_{nm} (Section 4.3.3) are also measured at the origin. Hence, to avoid spatial anomaly, we apply the spectral filter on the measured signal at the origin instead of at beamformer's output.

5.5.2 The proposed method

The odd spherical harmonics[1] are invisible on the XY plane due to the properties of the associated Legendre polynomials. Hence, a planar array is capable of extracting only the even soundfield coefficients. For an N^{th}-order soundfield, there exist $\big((N+1)(N+2)/2\big)$ active even modes, hence, the necessary condition to solve (5.1) with a planar array is

$$\left(\frac{(N+1)(N+2)}{2}\right)^2 \geq L + (V+1)^2 + 1. \tag{5.8}$$

Theoretically, it is possible to solve (5.1) using an arbitrary microphone array, provided that the criterion in (5.8) is satisfied. However, we consider uniformly distributed microphones in a circular array for our experiments which offer a simplified array structure with efficient computation techniques. Furthermore, such a circular array is common in various existing commercial audio appliances creating an opportunity for a seamless integration of the proposed technique.

To achieve the second design criteria, i.e. eliminating the beamformer, we place an additional microphone at the origin. Feeding the signal at origin to the Wiener filter's input agrees with definition of $\Phi_\ell(k)$ and $\Phi_r(k)$. Therefore, the estimated source signal under the new model becomes

$$\widehat{S}_\ell(k) = P(\boldsymbol{x}_0, k)\ \frac{\Phi_\ell(k)}{\sum\limits_{\ell'=1}^{L} \Phi_{\ell'}(k) + \Phi_r(k)} \tag{5.9}$$

where $\boldsymbol{x}_0 = (0,0,0)$ indicates the origin.

5.5.3 Extract the even coefficients using the proposed array structure

The extraction of the even soundfield coefficients using multiple circular arrays was first proposed in [68]. Here, we utilise a circular array with an additional microphone at the origin. We can readily calculate $\alpha_{00}(k)$ from the received signal

[1]The odd and even coefficients are decided based on the value of the corresponding $(n+|m|)$.

at the origin by setting $q = n = m = 0$ in (4.8) as

$$\alpha_{00}(k) = \sqrt{4\pi} P(\boldsymbol{x}_0, k). \tag{5.10}$$

Assuming that the circular array has a radius of R and contains Q omni-directional microphones, we obtain the sound pressure at each microphone using (4.8) by

$$P(\boldsymbol{x}_q, k) = \sum_{nm}^{N} \alpha_{nm}(k)\ j_n(kR)\ Y_{nm}(\frac{\pi}{2}, \phi_q) \tag{5.11}$$

where $N = \lceil kR \rceil$ and $q \in [1, Q]$. From the definition of the spherical harmonics in (2.13), we know

$$Y_{nm}(\frac{\pi}{2}, \cdot) = \begin{cases} \frac{1}{\sqrt{4\pi}}, & \text{if } n = 0 \\ 0, & \text{if } (n + |m|) \text{ is odd} \\ Y_{nm}(\frac{\pi}{2}, \cdot), & \text{otherwise.} \end{cases} \tag{5.12}$$

Hence, using (5.10) and (5.12) in (5.11), we obtain

$$P(\boldsymbol{x}_q, k) - P(\boldsymbol{x}_0, k)\ j_0(kR) = \sum_{\substack{nm \\ n \neq 0 \\ n+|m| \text{ even}}}^{N} \overline{\alpha}_{nm}(k)\ j_n(kR)\ Y_{nm}(\frac{\pi}{2}, \phi_q) \tag{5.13}$$

where $\overline{\alpha}_{nm}(k) = \big\{\alpha_{nm}(k) : n > 0;\ \text{and } (n + |m|) \text{ is even}\big\}$. Considering all the microphones on the circular array, we write (5.13) in a matrix form as

$$\begin{bmatrix} \overline{P}_1 \\ \vdots \\ \vdots \\ \overline{P}_Q \end{bmatrix} = \begin{bmatrix} \Lambda_{1-1}(\phi_1) & \dots & \Lambda_{NN}(\phi_1) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \Lambda_{1-1}(\phi_Q) & \dots & \Lambda_{NN}(\phi_Q) \end{bmatrix} \begin{bmatrix} \overline{\alpha}_{1-1} \\ \vdots \\ \vdots \\ \overline{\alpha}_{NN} \end{bmatrix} \tag{5.14}$$

where

$$\overline{P}_q = P(\boldsymbol{x}_q, k) - P(\boldsymbol{x}_0, k)\ j_0(kR) \tag{5.15}$$

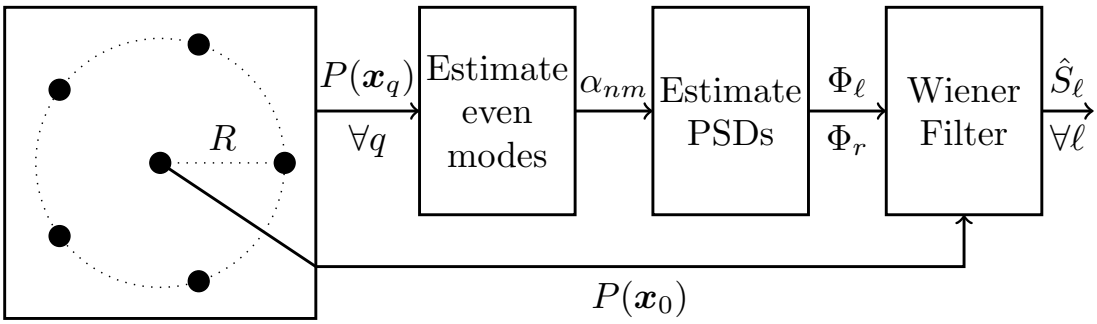


Figure 5.9: Block diagram of the proposed method using a planar array with 6 omni-directional microphones. FFT blocks and k-dependency are omitted for brevity.

and

$$\Lambda_{nm}(\phi_q) = j_n(kR)\ Y_{nm}(\frac{\pi}{2}, \phi_q). \tag{5.16}$$

The dependency on k is omitted in (5.14) for brevity. Note that, the right-most vector of (5.14) contains $\big((N+1)(N+2)/2 - 1\big)$ elements, hence, (5.14) can be solved for all $\overline{\alpha}_{nm}(k)$ as long as

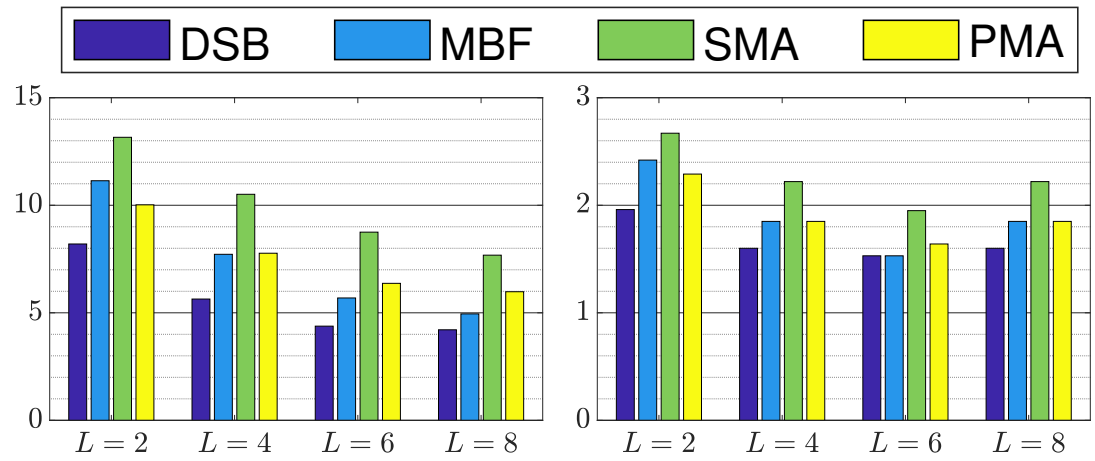
$$Q \geq \frac{(N+1)(N+2)}{2} - 1. \tag{5.17}$$

5.5.4 PSD estimation and source separation

Once we estimate all the even modes $\big[\alpha_{00}(k), \overline{\alpha}_{nm}(k)\big]$ using (5.10) and (5.14), we construct and solve (5.1) considering the even modes only, subject to the constraint mentioned in (5.8). Finally, we employ the single-channel Wiener filter of (5.9) to reconstruct each source signal separately. Fig. 5.9 shows the block diagram of the proposed method with the planar array structure.

5.6 Performance Evaluation with a Planar Array

We evaluated source separation performance using the proposed planar array through practical experiments as well as computer simulations. We considered $N = 2$ for this purpose and excluded the background noise in the modal coherence model for



(a) FWSegSNR (dB)

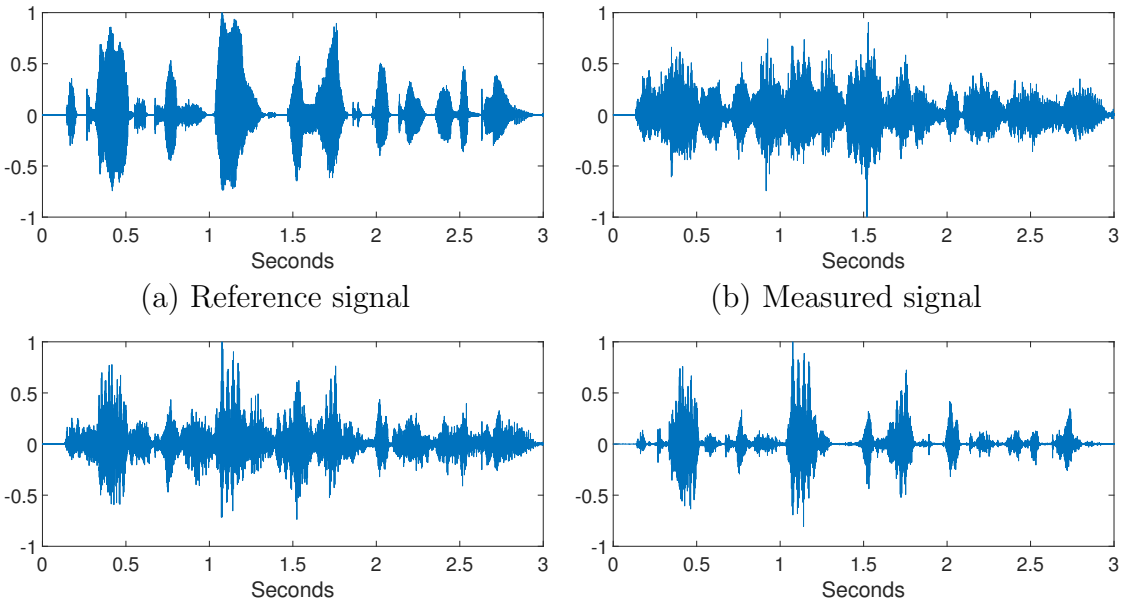
(b) PESQ

Figure 5.10: Average performances of the competing methods for non-reverberant cases.

tractability. Data processing was performed in the frequency domain with a 8 ms Hanning window, 50% frame overlap, a 128-point fast Fourier transform (FFT), and 8 kHz sampling frequency. The source directions were estimated using a spherical harmonics-based frequency-smoothed MUSIC algorithm [120]. All the sources were considered to be either above or below the XY-plane, which can be easily ensured with a proper placement of the array. The performance was measured through FWSegSNR and PESQ, as in the last experimental section. Each of the experiments was performed 20 times with mixed-gender random speech signals and the average values of the objective metrics are presented in the subsequent sections.

5.6.1 Non-reverberant case

We first consider a non-reverberant case in a simulated environment with $L = \{2, 4, 6, 8\}$ sources at random locations. We used the proposed planar array with $Q = 5$ and $R = 2$ cm. Fig. 5.10 compares the performance of the spatial coherence-based source separation using the proposed planar array (denoted as "PMA") as well as a 32-channel spherical microphone array (denoted as "SMA") of type *Eigenmike* with a conventional delay and sum beamformer (denoted as "DSB") and a



(a) Reference signal (b) Measured signal

(c)"DSB" output (d) "PMA" output

Figure 5.11: Estimated signal waveform of the first speaker in a 4-speaker non-reverberant environment.

multiple beamformer-based method [8] (denoted as "MBF"). For a fair comparison, we used the same number of microphones for the "DSB" and "MBF" methods as we used for the proposed method. The strong performance of "SMA" in Fig. 5.10 is expected as it was able to use the full coherence matrix due to the array structure and additional microphones. The proposed planar array used only the even modes to act on a partial coherence matrix, but still performed better compared to the conventional DSB. The performance comparison with "MBF" reveals that the proposed method exhibits better results in all of the cases except for $L = 2$ where advantage of "PMA" was concealed by the spatial aliasing error. Note that, "MBF" is designed to perform well when the number of beamformer is low [8], as is the case with $L = 2$. Hence, the relative performance gain achieved with the proposed method over "MBF" improved as the number of sound sources increases.

The estimated waveform for the first speaker in a 4-speaker system is shown in Fig. 5.11 which exhibits a good resemblance with the reference.

Table 5.2: Average performance in a practical reverberant room with 2 speakers.

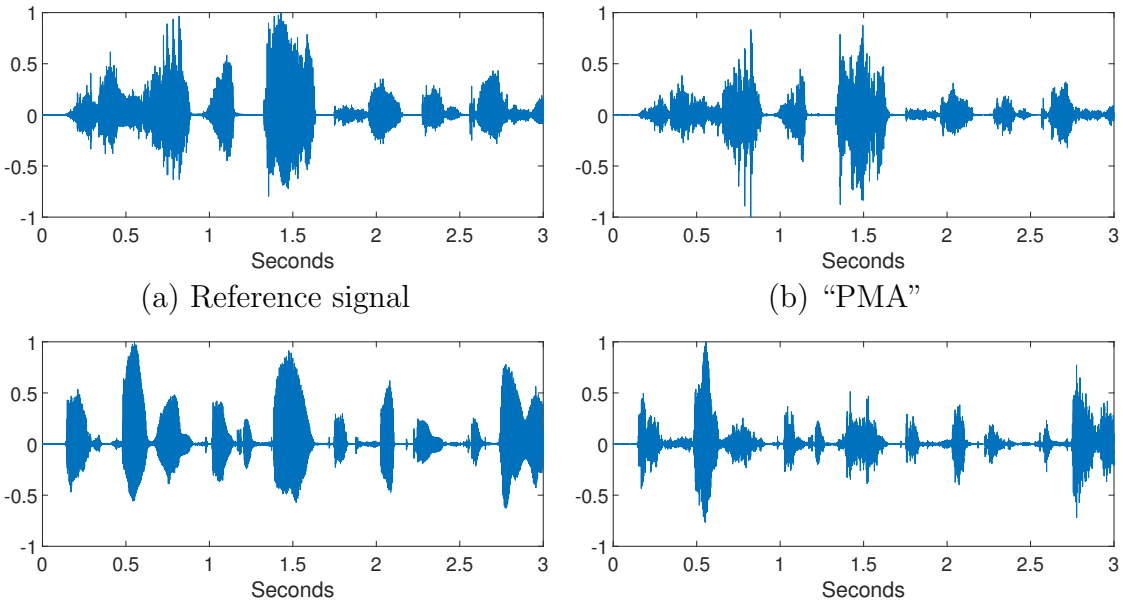
Metric	DSB	MBF	DMA	PMA
PESQ	1.94	1.90	2.2	2.22
FWSegSNR	3.57	3.79	4.45	3.97

5.6.2 Reverberant case

We also evaluated the performance in a practical room environment with reverberation. It is worth noting that, as a trade-off for using a small number of microphones on a single plane, we need to restrict the order of the reverberant soundfield power to $V \leq 1$ to avoid an under-determined system. The exclusion of the higher order reverberant soundfield power can introduce some artefacts at the final output, however, the contribution of the higher order modes to the total reverberant power is expected to be less prominent compared to the contribution of the lower order modes.

For the experimental validation, we used a planar array with $Q = 5$ and $R = 3$ cm. We compared the performance of the planar using a 16-channel hybrid differential microphone array [66] (denoted as "DMA") as well as with the "DSB" and "MBF" techniques referred in Section 5.6.1. The results with 2 sound sources are shown in Table 5.2 which suggest that the proposed method performs better compared to "DSB" and "MBF", and maintain a comparable performance with "DMA" despite having fewer number of microphones in the array. Note that, to achieve a better performance with a planar array based on the proposed algorithm, the planar array structure can be extended with multiple circles, such as [68], to extract the higher order modes.

Finally, we include the time domain waveforms of "PMA" output for the 2-speaker system in Fig. 5.12 to offer a visual comparison.



(a) Reference signal (b) "PMA"

(c) Reference signal (d) "PMA"

Figure 5.12: Estimated signal waveforms using the proposed planar array in a practical reverberant room with 2 speakers. (a)-(b) represent the first speaker while (c)-(d) shows the waveforms of the second speaker.

5.7 Summary

The objective of this chapter was to demonstrate a practical application of the modal coherence model. We performed source separation in different acoustic environments and measured the performance using a spherical and a planar microphone array. Below are the major outcomes from this chapter:

- We presented a complete performance overview of a source separation algorithm utilising the modal coherence model of a soundfield. We compared the performance of the proposed method with various conventional and contemporary techniques using a real-world dataset and in 7 different reverberant environments. The proposed method was found to outperform the competing methods in terms of audio quality and perception.

- The scalability of the modal coherence model is exploited to propose a source separation algorithm based on a partial coherence matrix. It allowed to em-

ploy a simpler planar array to achieve acceptable performance in reverberant and non-reverberant cases. The deployment of a planar array is beneficial from different practical aspects such as computational and manufacturing costs, processing time and seamless integration.

- We analysed the impact of different frequencies and soundfield orders on the outcome of the proposed algorithm.

This chapter studied different aspects of practical application of the modal coherence model we developed in the last chapter. We have noticed that both PSD estimation and source separation required the knowledge of source DOAs. So far, we have used the traditional methods to estimate the source locations. In the next chapter, we exploit the unique directional pattern of the modal coherence model to develop a better DOA estimation technique.

5.8 Related Publications

- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "PSD Estimation and Source Separation in a Noisy Reverberant Environment using a Spherical Microphone Array", IEEE/ACM Transactions on Audio Speech and Language Processing, Volume 26, Issue 9, pp. 1594–1607, 2018.

- **A. Fahim**, P. N. Samarasinghe, T. D. Abhayapala, and H. Chen, "A planar microphone array for spatial coherence-based source separation", Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6, Vancouver, Canada, August 2018.

This page intentionally left blank.

Chapter 6

Multi-Source DOA Estimation through Pattern Recognition of the Modal Coherence of a Reverberant Soundfield

Direction of arrival (DOA) estimation is an important prerequisite in soundfield and source separation. The traditional approaches to DOA estimation are known to suffer performance issues under strong reverberation and noisy environment. Hence, we pursue a data-driven approach to train a convolutional neural network to learn the unique directional pattern of the modal coherence of a soundfield we developed in the preceding chapters. Furthermore, we introduce a novel strategy of multi-source DOA estimation for overlapping sources that uses only single-source scenario during training. The proposed model achieves better performance and resource efficiency compared to the state-of-the-art methods in the same domain.

6.1 Introduction

We propose a novel multi-source direction of arrival (DOA) estimation technique using a convolutional neural network (CNN) algorithm which learns the modal coherence patterns of an incident soundfield through measured spherical harmonic coefficients. The data-driven approach allows the model to learn the evolution of acoustic environments and predict accurately in unknown conditions. The traditional approaches such as MUSIC and ESPRIT [12], [13] are known to be susceptible to strong reverberation and background noise [14]. Furthermore, they require to scan the whole DOA range at the run-time which affects latency of the system. Conversely, the beamforming-based approaches [15] experience degradation in their performance for closely-spaced sources due to the limitation of the spatial resolution. Recently, deep neural networks (DNN) are being used for DOA estimation [150], [151] to overcome the aforementioned limitations of the parametric domain. However, the computational complexity of the existing methods increases considerably as the number of sources and the DOA range increase.

In this work, we aim to improve the performance of DOA estimation in challenging acoustic environments by taking a CNN-based approach as well as propose a novel strategy to reduce the resource requirements and computational load. We train our model for individual time-frequency bins in the short-time Fourier transform spectrum by analysing the unique snapshot of modal coherence for each desired direction. The proposed method is capable of estimating simultaneously active multiple sound sources on a 3D space using a single-source training scheme. This single-source training scheme reduces the training time and resource requirements as well as allows the reuse of the same trained model for different multi-source combinations. The method is evaluated against various simulated and practical noisy and reverberant environments with varying acoustic criteria and found to outperform the baseline methods in terms of DOA estimation accuracy. Furthermore, as the training stage of the proposed algorithm involves single-source scenarios only, we can independently train our model for azimuths and elevations based on the same input dataset provided that we measure the soundfield for various source positions in each intended azimuth and elevation planes. This significantly improves the training efficiency of the joint azimuth and elevation estimation without

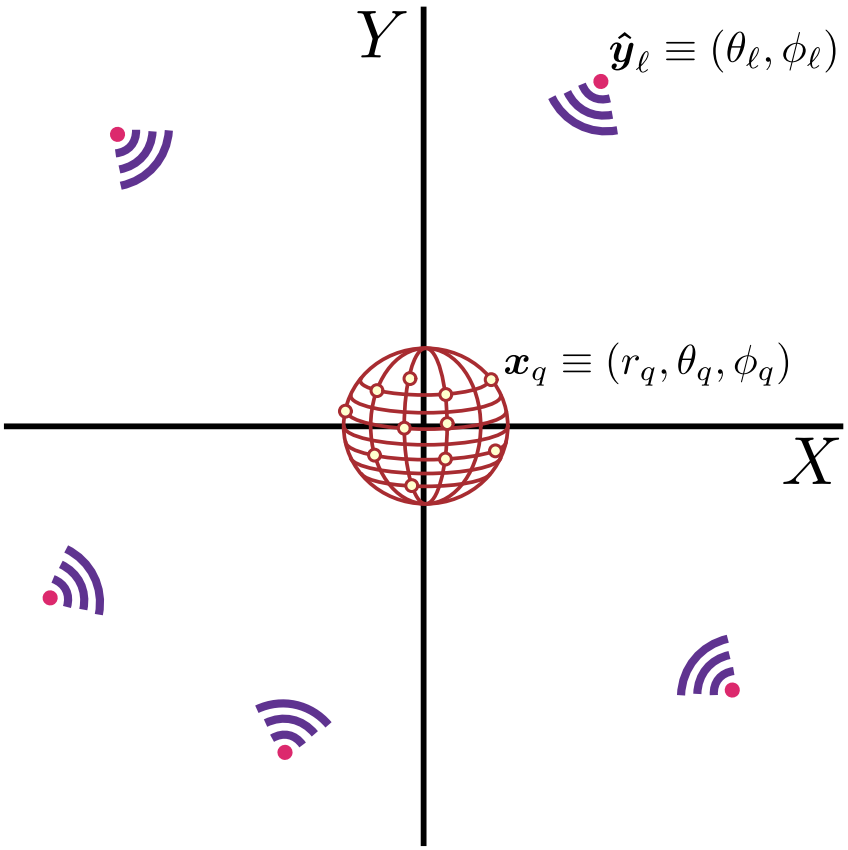


Figure 6.1: A graphical impression of a spherical microphone array setup in the presence of multiple sound sources. Array shape may differ depending on the spherical harmonic decomposition technique.

affecting the overall estimation accuracy.

The remainder of the chapter is structured as follows. Section 6.2 contains the problem statement and defines the objective of the work. In Section 6.3, we present a detailed description of the proposed model including different aspects of feature selection. Finally, in Section 6.4, we evaluate and analyse the performance of the proposed algorithm and compare it with a contemporary method based on objective metrics and graphical aid.

6.2 Problem Formulation

Consider L sound sources concurrently emitting sound in a reverberant room. The sound pressure observed by an omnidirectional microphone placed at a coordinate $\boldsymbol{x}_q \equiv (r_q, \theta_q, \phi_q)$ inside the room, where r_q, θ_q, and ϕ_q are the radius, elevation, and azimuth of point $\boldsymbol{x}_q$ in the spherical coordinate system, respectively, is expressed by

$$p(\boldsymbol{x}_q, t) = \sum_{\ell=1}^{L} h_\ell(\boldsymbol{x}_q, t) * s_\ell(t) \tag{6.1}$$

where t is the discrete time index, $h_\ell(\boldsymbol{x}_q, t)$ is the room impulse response (RIR) between the ℓ^{th} source position and $\boldsymbol{x}_q$, $s_\ell(t)$ is the ℓ^{th} source signal, and $*$ denotes the convolution operation. The corresponding frequency domain representation of (6.1) in short-time Fourier transform (STFT) domain can be obtained using the multiplicative model of convolution and is formulated as

$$P(\boldsymbol{x}_q, \tau, k) = \sum_{\ell=1}^{L} S_\ell(\tau, k) H_\ell(\boldsymbol{x}_q, k) \tag{6.2}$$

where $\{P, S, H\}$ represent the corresponding signals of $\{p, s, h\}$ in the STFT domain, τ is the time frame index, $k = 2\pi f/c$, f denotes the frequency, and c is the speed of sound propagation. Henceforth, τ is omitted for brevity as we shall treat each of the time frames independently.

In this work, we intend to estimate the individual DOAs in the presence of multiple concurrent sound sources, i.e., we want to estimate $\widehat{\widehat{\boldsymbol{y}}}_\ell \equiv (\widehat{\theta}_\ell, \widehat{\phi}_\ell)\ \forall \ell \in [1, L]$, given a set of measured sound pressure $p(\boldsymbol{x}_q, t)\ \forall q \in [1, Q]$ or the corresponding spherical harmonic coefficients[1] of a mixed soundfield. We pose the DOA estimation as a CNN classification problem where we sample the intended DOA range into discrete sets $\Theta = \left\{\theta_a\right\}_{a\in[1,I]}$ for elevations and $\Phi = \left\{\phi_b\right\}_{b\in[1,J]}$ for azimuths. Thereafter, we propose a feature unique to each angle and train a CNN framework individually for each of the members of Θ and Φ. Finally, during the evaluation, the CNN model finds the closest match of the true DOA $\widehat{\boldsymbol{y}}_\ell \equiv (\theta_\ell, \phi_\ell)\ \forall \ell$ in the DOA sets Θ and Φ based on its learning and accurately combines the independent

[1]The spherical harmonic decomposition technique is described in Section 4.3.2.

estimations $\widehat{\theta}_\ell$ and $\widehat{\phi}_\ell$ for each individual source to achieve full DOA estimation.

6.3 CNN-based DOA Estimation

CNN is a popular technique in the deep learning domain, and is predominantly used in computer vision applications. The input, often a 2D or 3D tensor, goes through multiple convolution filters followed by a traditional fully-connected neural network. In this work, we pose the DOA estimation problem as an image-classification problem where the input image represents the modal coherence of the soundfield.

6.3.1 Modal Framework

We are going to construct a feature utilising the concept of modal coherence of a reverberant soundfield in order to efficiently train a CNN. In Chapter 4, we developed a closed-form expression of the modal coherence of a reverberant soundfield as

$$\mathbb{E}\Big\{\alpha_{nm}(k)\alpha^*_{n'm'}(k)\Big\} = \sum_{\ell=1}^{L}\mathbb{E}\Big\{\Big|S_\ell(k)\Big|^2\Big\}\Bigg(\mathbb{E}\Big\{\Big|G^{(d)}_\ell(k)\Big|^2\Big\}\ \Upsilon^{n'm'}_{nm}(\widehat{\boldsymbol{y}}_\ell)+ \sum_{vu}^{V}\mathbb{E}\Big\{\gamma^{(\ell)}_{vu}(k)\Big\}\ \Psi^{m,m',u}_{n,n',v}\Bigg) \tag{6.3}$$

where $\mathbb{E}\{\cdot\}$ denotes expected value and

$$\sum_{vu}^{V}\ \mathbb{E}\Big\{\gamma^{(\ell)}_{vu}(k)\Big\}\ Y_{vu}(\widehat{\boldsymbol{y}}) = \mathbb{E}\Big\{\Big|G^{(r)}_\ell(k,\widehat{\boldsymbol{y}})\Big|^2\Big\} \tag{6.4}$$

$$\Upsilon^{n'm'}_{nm}(\widehat{\boldsymbol{y}}_\ell) = C_{nn'}\ Y^*_{nm}(\widehat{\boldsymbol{y}}_\ell)\ Y_{n'm'}(\widehat{\boldsymbol{y}}_\ell) \tag{6.5}$$

$$\Psi^{m,m',u}_{n,n',v} = C_{nn'}\ W^{u,m,m'}_{v,n,n'} \tag{6.6}$$

$$C_{nn'} = 16\pi^2 i^{n-n'} \tag{6.7}$$

$$W_{v,n,n'}^{u,m,m'} = (-1)^m \sqrt{\frac{(2v+1)(2n+1)(2n'+1)}{4\pi}} \begin{pmatrix} v & n & n' \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v & n & n' \\ u & -m & m' \end{pmatrix} \tag{6.8}$$

where $(\cdot)$ in (6.8) represents Wigner-3j symbol [189]. For temporal processing, it is common to estimate the expected value by applying the exponential moving average technique on the instantaneous measurements, i.e.,

$$\mathbb{E}\Big\{\alpha_{nm}(\tau,k)\alpha^*_{n'm'}(\tau,k)\Big\} = \beta\ \mathbb{E}\Big\{\alpha_{nm}(\tau-1,k)\alpha^*_{n'm'}(\tau-1,k)\Big\} + (1-\beta)\ \alpha_{nm}(\tau,k)\alpha^*_{n'm'}(\tau,k) \tag{6.9}$$

where $\beta \in [0,1]$ is a smoothing factor.

6.3.2 Feature selection

Intuitively, the soundfield coefficients α_{nm} work as natural beamformers in the modal domain due to the inherent properties of the spherical harmonic functions. Hence, the energy distribution of α_{nm} among different modes can be used as a clue for understanding the source directionality. However, there only exists a limited number of active modes in the low frequencies which might prove insufficient to train a neural network for high spatial resolution scenario, especially in a reverberant environment. Therefore, we use the modal coherence model of (6.3) to construct our input feature. For a multi-source scenario, it is common to assume W-disjoint orthogonality [16] in the STFT domain, i.e., only a single sound source remains active in each time-frequency (TF) bin of the STFT spectrum. Under the W-disjoint orthogonality assumption, (6.3) takes the following form

$$\mathbb{E}\Big\{\alpha_{nm}(k)\alpha^*_{n'm'}(k)\Big\} = \mathbb{E}\Big\{\Big|S_{\ell'}(k)\Big|^2\Big\}\Bigg(\mathbb{E}\Big\{\Big|G^{(d)}_{\ell'}(k)\Big|^2\Big\}\ \Upsilon^{n'm'}_{nm}(\widehat{\boldsymbol{y}}_{\ell'}) + \sum_{vu}^{V} \mathbb{E}\Big\{\gamma^{(\ell')}_{vu}(k)\Big\}\ \Psi^{m,m',u}_{n,n',v}\Bigg) \tag{6.10}$$

where $\ell' \in [1, L]$. Note that, (6.10) remains true for a single-source scenario as well.

For audio signals with variable spectral densities, e.g., speech signals, it is intu-

itive to select a feature based on the relative transfer function to make the feature independent to the variations in the input signal [206]. Following a similar reasoning, we define the relative modal coherence (RMC) as

$$\frac{\mathbb{E}\Big\{\alpha_{nm}(k)\alpha^{*}_{n'm'}(k)\Big\}}{\mathbb{E}\Big\{\alpha_{00}(k)\alpha^{*}_{00}(k)\Big\}} = \frac{\mathbb{E}\Big\{\Big|G^{(d)}_{\ell'}(k)\Big|^2\Big\}\ \Upsilon^{n'm'}_{nm}(\widehat{\boldsymbol{y}}_{\ell'}) + \sum_{vu}^{V}\mathbb{E}\Big\{\gamma^{(\ell')}_{vu}(k)\Big\}\ \Psi^{m,m',u}_{n,n',v}}{\mathbb{E}\Big\{\Big|G^{(d)}_{\ell'}(k)\Big|^2\Big\}\ \Upsilon^{00}_{00}(\widehat{\boldsymbol{x}}_{\ell'}) + \sum_{vu}^{V}\mathbb{E}\Big\{\gamma^{(\ell')}_{vu}(k)\Big\}\ \Psi^{0,0,u}_{0,0,v}} \tag{6.11}$$

$$= \frac{\mathbb{E}\Big\{\Big|G^{(d)}_{\ell'}(k)\Big|^2\Big\}\ \Upsilon^{n'm'}_{nm}(\widehat{\boldsymbol{x}}_{\ell'}) + \sum_{vu}^{V}\mathbb{E}\Big\{\gamma^{(\ell')}_{vu}(k)\Big\}\ \Psi^{m,m',u}_{n,n',v}}{4\pi\ \mathbb{E}\Big\{\Big|G^{(d)}_{\ell'}(k)\Big|^2\Big\} + \frac{16\pi^2}{\sqrt{4\pi}}\ \mathbb{E}\Big\{\gamma^{(\ell')}_{00}(k)\Big\}} \tag{6.12}$$

where (6.12) is derived using (6.5) - (6.8) in (6.11). From (6.12) it is evident that the relative modal coherence has a direct relation with the source position in a particular room. However, with multiple active sources in a strong reverberant environment, (6.12) introduces additional complexity due to the additive terms in the denominator. On the other hand, the modal coherence of (6.10) offers a simpler alternative to train a CNN due to the fact that the mode-independent term $\Big\{\Big|S_{\ell'}(k)\Big|^2\Big\}$ of (6.10) acts merely as a constant scaling factor across different TF bins without altering the relative strength between different modes inside a TF bin. The use of (6.10) as a feature reduces the complexity by eliminating location-dependency from the denominator compared to RMC. Hence, we pose the DOA estimation problem as an image-identification problem for CNN where the feature snapshot is defined as the modal coherence of the soundfield in the individual TF bins of the STFT spectrum

$$\widehat{\boldsymbol{\mathcal{F}}}_{\mathrm{mc}}(k) = \Big\{\mathbb{E}\Big\{\alpha_{nm}(k)\alpha^{*}_{n'm'}(k)\Big\} : n \in [0,N], m \in [-n,n], n' \in [0,N], m' \in [-n',n']\Big\} \tag{6.13}$$

where $\widehat{\boldsymbol{\mathcal{F}}}_{\mathrm{mc}}$ is considered as an image consisting of $[\mathcal{N}\times\mathcal{N}]$ complex-valued pixels with $\mathcal{N} = (N+1)^2$ is the total number of modes. Note that, $\widehat{\boldsymbol{\mathcal{F}}}_{\mathrm{mc}}$ is a frequency-dependent function due to the frequency dependency of α_{nm}, hence, we need to

collect $\widehat{\boldsymbol{\mathcal{F}}}_{\text{mc}}$ from different frequency bands for training so that the model can learn the frequency variations of the feature for the same source position. This deviation is analogous to the transformed image conundrum in an image-identification problem.

We also need to train the CNN model for various amplification levels due to the presence of the source PSD term $\mathbb{E}\left\{\left|S_{\ell'}(k)\right|^2\right\}$ in $\widehat{\boldsymbol{\mathcal{F}}}_{\text{mc}}$ (analogous to train a neural network to accommodate the differences in brightness of the same image). This can be achieved through training the CNN model with any non-white random audio signal such that the signal has a variable PSD in both time and frequency directions of the STFT spectrum.

Finally, since a CNN model is best suited to work with real data, we convert our 2D complex-valued feature $\widehat{\boldsymbol{\mathcal{F}}}_{\text{mc}}$ into corresponding 3D tensor $\boldsymbol{\mathcal{F}}_{\text{mc}}$ of $[\mathcal{N} \times \mathcal{N} \times 2]$ dimension such that

$$\boldsymbol{\mathcal{F}}_{\text{mc}} = \left\langle \left\langle \mathcal{R}\{\widehat{\boldsymbol{\mathcal{F}}}_{\text{mc}}\},\ \mathcal{I}\{\widehat{\boldsymbol{\mathcal{F}}}_{\text{mc}}\} \right\rangle \right\rangle_3 \tag{6.14}$$

where $\langle\langle \cdot,\ \cdot \rangle\rangle_3$ stacks two matrices in the 3^{rd} dimension and $\mathcal{R}\{\cdot\}$ and $\mathcal{I}\{\cdot\}$ denote the real and imaginary part, respectively.

Fig. 6.2 shows the normalised snapshots of $\boldsymbol{\mathcal{F}}_{\text{mc}}$ captured at random time instants at 1500 Hz. For a CNN model to work with our input features, we want them to be time-independent for the same source position in a room irrespective of the nature of the audio signal. Indeed, as we observe from Fig. 6.2, $\boldsymbol{\mathcal{F}}_{\text{mc}}$ changes as a function of source angle and remains fairly constant across time.

6.3.3 TF bin processing

During both training and evaluation phases, the proposed CNN framework processes each TF bin independently, i.e., it learns the directional patterns based on the spatial distribution of the TF bin energy. Hence, it is important to consider only the TF bins with a significant energy level to avoid misleading the neural network. However, due to the sparse nature of speech signals in both time and frequency, a large proportion of the TF bins usually ends up having low energy. The sparsity in time can be addressed with a suitably designed voice activity detector,

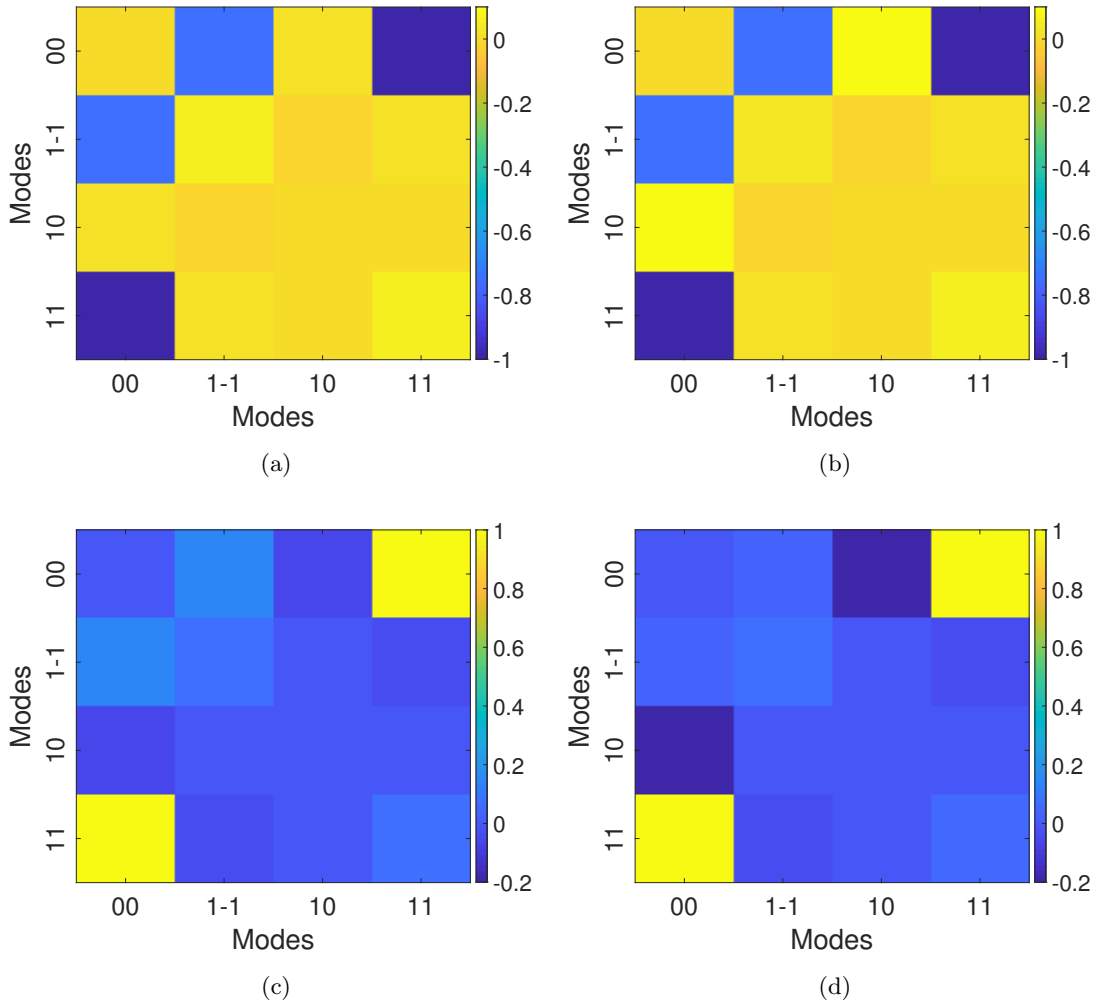


Figure 6.2: Normalised $\boldsymbol{\mathcal{F}}_{\text{mc}}$ at different time instants. The snapshot is taken at 1500 Hz with a speech source present at (a)-(b) $(\theta, \phi) = (60°, 60°)$ and (c)-(d) $(\theta, \phi) = (60°, 120°)$.

but the sparsity along the frequency can still mislead the CNN. Hence, to exclude the low-energy TF bins from the training and evaluation datasets, a energy-based pre-selection of TF bins is required where we drop all the TF bins with an average energy below a certain threshold. If $\mathcal{T}_{\text{all}}$ is denoted as the collection of all the TF bins, we can define a new set $\mathcal{T}_{\text{act}} \subseteq \mathcal{T}_{\text{all}}$ such that

$$\mathcal{T}_{\text{act}} = \{\kappa \in \mathcal{T}_{\text{all}} : E_{\kappa} \geq E_{\text{min}}\} \tag{6.15}$$

where E_κ is the average energy of the spatial coherence matrix for the κ^{th} TF bin and $E_{\min}$ is the minimum energy threshold. This lowest energy threshold can be a preset based on empirical measurements, or it can be set dynamically based on the average energy of all the TF bins in the processing block. However, the average energy of the processing block can be low when the number of low-energy TF bins is high. We can also set the minimum energy level at the $\mathcal{K}^{th}$ percentile of the average energy distribution of all TF bins, where $\mathcal{K}$ is usually large for speech signal, as long as we are able to make at least a high-level prediction about the energy distribution among the TF bins. Note that, (6.15) should be applied at both training and evaluation stage, however, $E_{\min}$ does not need to be the same.

A second issue may arise when a TF bin violates the W-disjoint orthogonality principle. In the proposed algorithm, CNN predicts the most dominant source in each TF bin which are later combined in a clustered histogram to reach a global outcome. However, as shown in [16], the number of TF bins violating the W-disjoint orthogonality increases as the number of simultaneous sources increases. When a TF bin contains significant energy from multiple sources, the prediction of the CNN model can be arbitrary. To circumvent the uncertainty in prediction due to the violation of W-disjoint orthogonality principle, we only consider the predictions in the TF bins where the CNN model predicts a single DOA with a high confidence level. Hence, if we define the probability score for each TF bin as

$$\boldsymbol{\mathcal{P}}_{\kappa,\Theta} = \left\{\mathcal{P}_\kappa(\theta)\right\}_{\theta\in\Theta} \tag{6.16}$$

$$\boldsymbol{\mathcal{P}}_{\kappa,\Phi} = \left\{\mathcal{P}_\kappa(\phi)\right\}_{\phi\in\Phi} \tag{6.17}$$

where $\mathcal{P}_\kappa(\theta)$ and $\mathcal{P}_\kappa(\phi)$ are the CNN's predicted score for the corresponding elevation and azimuth classes at the κ^{th} TF bin, the final DOA estimation at the evaluation stage should be based on the set $\mathcal{T}_{\text{test}} \subseteq \mathcal{T}_{\text{act}}$ such that

$$\mathcal{T}_{\text{test}} = \left\{\kappa \in \mathcal{T}_{\text{act}} : \max\left\{\boldsymbol{\mathcal{P}}_{\kappa,\Theta}\right\} \geq \mathcal{P}_{\min} \text{ and } \max\left\{\boldsymbol{\mathcal{P}}_{\kappa,\Phi}\right\} \geq \mathcal{P}_{\min}\right\} \tag{6.18}$$

where $\mathcal{P}_{\min}$ denotes the minimum confidence level.

6.3.4 CNN architecture

We utilise a CNN to estimate source DOA based on the local connectivity of the modal coherence coefficients. A CNN topology typically consists of multiple convolution layers followed by fully-connected networks. For DOA estimation, we perform multi-output multi-class classification where we share the same convolution layer structure to predict both the azimuth and elevation using separate fully connected heads at the last stage - each responsible for predicting either azimuth or elevation. We opt for a classification-based approach due to the limited resolution of the practical dataset. However, the proposed technique can be extended for a regression-based model subject to the availability of a denser training grid to learn the evolution of dynamic reverberation characteristics.

In each convolutional layer, we use 64 spatial filters of 2×2 size to learn the spatial coherence pattern for each desired point in a predefined DOA grid. As our feature is defined as the modal coherence for each TF bin, it is important to consider 2D filters in the convolution layers. A rectified linear unit (ReLU) activation follows the convolution layer at each stage. The evaluation is done with 8 convolution layers with zero padding to keep the output size identical for all the layers. The final convolution layer is connected to 2 fully-connected layers that use ReLU activation. Finally, two separate fully connected heads responsible for azimuth and elevation estimation, respectively, are used with Sigmoid activation. A Sigmoid activation is chosen over Softmax in the last stage as it allows us to perform prediction-based TF bin selection to remove the bins with low confidence, as described in Section 6.3.3.

Due to the W-disjoint orthogonality assumption, ideally each TF bin is designated with a single DOA and can be classified using a multi-class classification network using a Softmax activation-based categorical cross-entropy loss. However, in a practical environment with multiple simultaneously active sources, it is unrealistic to expect each of the TF bins to honour the W-disjoint orthogonality. Therefore, it is possible to find occasional TF bins whose energy are contributed by multiple sound sources. Such a TF bin produces a feature snapshot which does not match with any of the patterns learned by the model during the single-source training stage. In such cases it is expected that the output will not have a large

prediction score for any of the classes. Hence, we use binary cross-entropy loss function with Sigmoid activation instead of categorical cross-entropy in order to independently predict the probability of each individual class in every TF bin. This approach allows us to enforce the criterion mentioned in (6.18).

Detailed parameter settings are included in the experimental results section.

6.3.5 Training the model

In the training stage, we train the model based on the feature snapshots in a single source scenario. Each training data is labelled independently for azimuths and elevations. The model is trained for the elevation set Θ and the azimuth set Φ in different azimuth and elevation planes, respectively. Once the model learns the patterns for each of the intended directions, we independently predict the elevations and azimuths for any number of concurrent sources as long as the W-disjoint orthogonality principle majorly holds. This is a more realistic approach than training the model for each possible angular combination [150], [152] which becomes a resource-intensive operation as the number of classes or the number of simultaneous sources increases. Furthermore, the proposed method does not require retraining the model every time an additional source appears in the mixture.

6.3.6 DOA estimation

First, we jointly pick the highest probable elevation and azimuth classes for each TF bin in $\mathcal{T}_{\text{test}}$ to form a prediction multiset $\mathcal{X}$ such that

$$\mathcal{X} = \Big\{ \Big(\arg\max_{\theta \in \Theta} \big\{ f : \theta \mapsto \mathcal{P}_{\kappa}(\theta) \big\},\ \arg\max_{\phi \in \Phi} \big\{ f : \phi \mapsto \mathcal{P}_{\kappa}(\phi) \big\} \Big) : \kappa \in \mathcal{T}_{\text{test}} \Big\}. \quad (6.19)$$

Let Ω be a set of all possible combinations of azimuths and elevation classes, i.e.,

$$\Omega = \{(\theta, \phi)\}_{\theta \in \Theta,\ \phi \in \Phi}. \quad (6.20)$$

Algorithm 2: Algorithm for DOA estimation - training stage

Data: $\Theta, \Phi, \alpha_{nm}(\theta, \phi)\ \forall nm, \forall\theta \in \Theta, \forall\phi \in \Phi$

1 Calculate spatial coherence $\mathbb{E}\Big\{\alpha_{nm}(k)\alpha^*_{n'm'}(k)\Big\}$ in each TF bin using (6.9);
2 Get $\widehat{\boldsymbol{\mathcal{F}}}_{\text{mc}}\ \forall\theta \in \Theta, \forall\phi \in \Phi$ using (6.13);
3 Apply (6.15) to filter out low energy TF bins and get $\mathcal{T}_{\text{act}}$;
4 Use $\mathcal{T}_{\text{act}}$ to train the model using the parameters in Table 6.1 independently for Θ and Φ;
5 Save the model

Subsequently, we define an accumulation function $\mathcal{Z}$ which calculates the number of repetitions of any element of Ω in $\mathcal{X}$,

$$\mathcal{Z} : \vartheta \in \Omega \mapsto \mathcal{M}_{\mathcal{X}}(\vartheta) \tag{6.21}$$

where $\mathcal{M}_{\mathcal{X}}(\vartheta)$ denotes the multiplicity of ϑ in the multiset $\mathcal{X}$, i.e., the number of times ϑ occurs in $\mathcal{X}$. Hence, the simplest way of multi-source DOA estimation is to pick L largest peaks in $\mathcal{Z}(\vartheta)$

$$\big\{\widehat{\widehat{\boldsymbol{x}}}_\ell\big\}_{\ell\in[1,L]} = \big\{\vartheta' \in \Omega : \mathcal{Z}(\vartheta') \text{ is one of L largest peaks in } \mathcal{Z}\big\}. \tag{6.22}$$

However, in case of a noisy prediction for a multi-source environment, the aforementioned technique can cause erroneous results. For example, in a 2-source environment, if the true DOA of the prominent source lies between two adjacent classes, both the adjacent classes for the prominent source might occur more frequently than the true class corresponding to the weaker source. To avoid such a scenario, a more robust technique is to apply a suitable clustering algorithm, such as k-means [207] or density-based [208] clustering, to divide $\mathcal{X}$ into L clusters and pick the peak in each cluster using

$$\widehat{\widehat{\boldsymbol{x}}}_\ell = \underset{\vartheta\in\Omega_\ell\subset\Omega}{\arg\max}\big\{\mathcal{Z}(\vartheta)\big\},\ \forall\ell \in [1, L] \tag{6.23}$$

where Ω_ℓ is a subset of Ω containing all points in ℓ^{th} cluster.

The training and evaluation steps are outlined in Algorithms 2 and 3, respectively.

Algorithm 3: Algorithm for DOA estimation - evaluation stage

Data: $\alpha_{nm}\ \forall nm$

Calculate spatial coherence $\mathbb{E}\Big\{\alpha_{nm}(k)\alpha^*_{n'm'}(k)\Big\}$ in each TF bin using (6.9);
Get $\widehat{\boldsymbol{\mathcal{F}}}_{\text{mc}}$ using (6.13);
Apply (6.15) to filter out low energy TF bins and get $\mathcal{T}_{\text{act}}$;
Calculate the probability of each classes in Θ and Φ for the TF bins in $\mathcal{T}_{\text{act}}$ using the model saved during training;
Apply (6.18) to get $\mathcal{T}_{\text{test}}$;
Apply (6.19) to form the prediction multiset $\mathcal{X}$;
if $L == 1$ **then**
    $\hat{\hat{\hat{\boldsymbol{x}}}} = \arg\max_{\vartheta\in\Omega} \big\{\mathcal{Z}(\vartheta)\big\}$;
else
    Using a suitable clustering algorithm, divide $\mathcal{X}$ into L clusters;
    Use (6.23) to estimate L source directions.;
end

6.4 Experimental Results and Discussion

In this section, we present the experimental results, comparison, and discussion of the proposed algorithm with the contemporary counterparts.

6.4.1 Experimental methodology

We evaluated the proposed method in simulated and practical environments under different room conditions. The parameter settings used in the evaluation are listed in Table 6.1. We assessed the performance of the model in 3 simulated room environments (room S1, S2, and S3 in Table 6.2) generated using a RIR Generator [209] as well as with the recordings from a practical room (room P1 in Table 6.2) in the presence of babble noise. The reverberation time (T_{60}) and direct to reverberation ratio (DRR) shown in Table 6.2 for room P1 were calculated using the techniques outlined in [200]. We only considered the first-order harmonic coefficients which need at least 4 microphones to calculate, however, in the result section, we used recordings from 9 microphones oriented on a spherical grid suggested in [210] for evaluating the proposed as well as the competing methods.

Table 6.1: Parameter settings for the experiments

Name	Value
Model parameters	
$\mathcal{P}_{\min}$	0.5
$E_{\min}$	Percentile-based
$\mathcal{K}$	90
N	1
CNN parameters	
Input size	$[4 \times 4 \times 2]$
# Convolution layers	8
# Conv. filters	64 ($[2 \times 2]$)
# Dense layers	2 (512)

We used random mixed-gender speech signals from the TIMIT corpus [201] to synthesise the reverberant signals from the measured/simulated RIRs. The spherical harmonic decomposition, as described in Section 4.3.2, was performed on the reverberant signals to calculate the spherical harmonic coefficients up to first order. Note that, the proposed method is independent of the type and shape of the sensor array as long as the array is capable of performing spherical harmonic decomposition[2].

The CNN architecture has been discussed and presented in Section 6.3.4 and Table 6.1. The implementation was done in Python using Keras [211] running on top of TensorFlow [212]. For the proposed method, the TF bin-level predictions were accumulated and clustered using k-means algorithm (step 10 in Algorithm 3) assuming that the number of active sources was known as *a priori*, however, certain algorithms offer to cluster the data without the advance knowledge of the number of sources [208]. Furthermore, as k-means algorithm works with the Euclidean geometry, we converted the predicted DOAs to corresponding Cartesian coordinates on a unit sphere before clustering the data.

The processing was done at 16 kHz sampling frequency. The STFT used a 16ms

[2]A number of alternate array structures are available in the literature for capturing spherical harmonic coefficients, a few can be found in [63], [65]–[68], [71].

Table 6.2: Test environments. d_{sm} denotes source to microphone distance.

Room	Dimension	T_{60}	DRR	d_{sm}
P1	$[11 \times 7.5 \times 2.75]$ m	640 ms	-0.6 dB	2.8 m
S1	$[6 \times 4 \times 3]$ m	200 ms	-	1 m
S2	$[7 \times 6 \times 3]$ m	300 ms	-	1 m
S3	$[8 \times 6 \times 3]$ m	500 ms	-	1 m

Hanning window, 50% overlap, and 256-point discrete Fourier transform (DFT). We only utilised the frequency range $500 - 2000$ Hz for DOA estimation. A 30s long speech was used to synthesise the data for training, however, the actual number of features reduced significantly after applying (6.15) to filter out low-energy TF bins. The majority of the results and discussions in this section are presented for azimuth estimation only considering the fact that the estimation of evaluation is independent of the azimuth estimation and follows the same mechanism. However, in Section 6.4.3, we have demonstrated how a joint azimuth and elevation estimation can be performed using the proposed method.

6.4.2 Baseline methods and evaluation metrics

The performance of the proposed algorithm is compared with a recent CNN-based DOA estimation method proposed in [150] (subsequently denoted as "CNN-PH") where it was already shown that "CNN-PH" outperforms conventional parametric methods like MUSIC and SRP-PHAT. For a fair comparison, we kept the CNN architecture and other evaluation criteria same in all possible ways. We used the same 9-microphone setup as described in Section 6.4.1 for the competing methods unless mentioned otherwise. The convolution filter size for "CNN-PH" was set to $[2 \times 1]$ as per the recommendation of the authors [150] whereas we applied $[2 \times 2]$ filters with the proposed method. The difference in filter size between the competing methods comes from the fact that the feature used in "CNN-PH" spans across the frequency band where multiple active sources can be present in the horizontal dimension. On the other hand, the proposed method uses the modal coherence snapshot of a single TF bin as a feature where only one active source is

expected due to the assumption of W-disjoint orthogonality.

To evaluate the performance, we first defined the prediction error for the ℓ^{th} source in a single test by the angular difference between the true and the estimated points at the origin of a unit sphere, i.e.,

$$\Delta_\ell = \cos^{-1}\Big[\cos(\widehat{\theta}_\ell)\cos(\theta_\ell) + \sin(\widehat{\theta}_\ell)\sin(\theta_\ell)\cos(\widehat{\phi}_\ell - \phi_\ell)\Big]. \tag{6.24}$$

As we are posing the DOA estimation as a classification problem, the mean error can be misleading unless the angular difference between adjacent classes are very small. Hence, instead we propose to use performance metrics based on estimation accuracy. At first, we define the multi-source DOA classification accuracy as the percentage of the correct predictions, i.e.,

$$\eta_{\text{acc}} = \frac{\mathcal{M}_{\{\{\Delta_\ell\}\}}(0)}{\overline{\overline{\{\{\Delta_\ell\}\}}}} \times 100\% \tag{6.25}$$

where $\{\{\Delta_\ell\}\}$ is a multiset containing $\Delta_\ell\ \forall \ell$ for all the tests and $\overline{\overline{\cdot}}$ denotes the cardinality of the underlying multiset. Note that, for a single test, the sequence of the true and estimated DOAs need not be in the same order, hence, we map them in such a way that $\mathcal{M}_{\{\{\Delta_\ell\}\}}(0)$ is maximised.

Occasionally, the definition of (6.25) may fail to offer the full picture as it does not take it into consideration how far a wrong prediction deviates from the true value although the adjacent classes are highly correlated in a DOA classification task. Hence, we define another accuracy metric, termed as adjacent accuracy, where we consider the predictions for adjacent classes as true positives as well, i.e.,

$$\eta_{\text{adj}} = \frac{\mathcal{M}_{\{\{\max[0,\Delta_\ell - \Delta_\Omega]\}\}}(0)}{\overline{\overline{\{\{\Delta_\ell\}\}}}} \times 100\% \tag{6.26}$$

where Δ_Ω is the angular separation between two adjacent classes. A high η_{adj} with low η_{acc} indicates that the transition of the feature pattern is not very sharp between the adjacent classes, a phenomenon expected in a noisy environment.

All the results presented in the subsequent sections are based on the accumulation of the results of 50 random experiments in each test case. Each experiment was evaluated with random source positions and subsequently added with random

Gaussian noise unless specified otherwise.

6.4.3 Results and discussions

In this section, we discuss and compare DOA estimation performances under different criteria and room environments. During the experiments, the microphone array was placed at the centre of the room at 1m height. For the proposed method, we completed the training once per room considering a single-source scenario and used the same trained model at the testing stage irrespective of the number of simultaneously active sources. A 30s long speech was synthesised during the training stage, however, we only used the top 10% STFT bins based on TF bin energy to train the network[3].

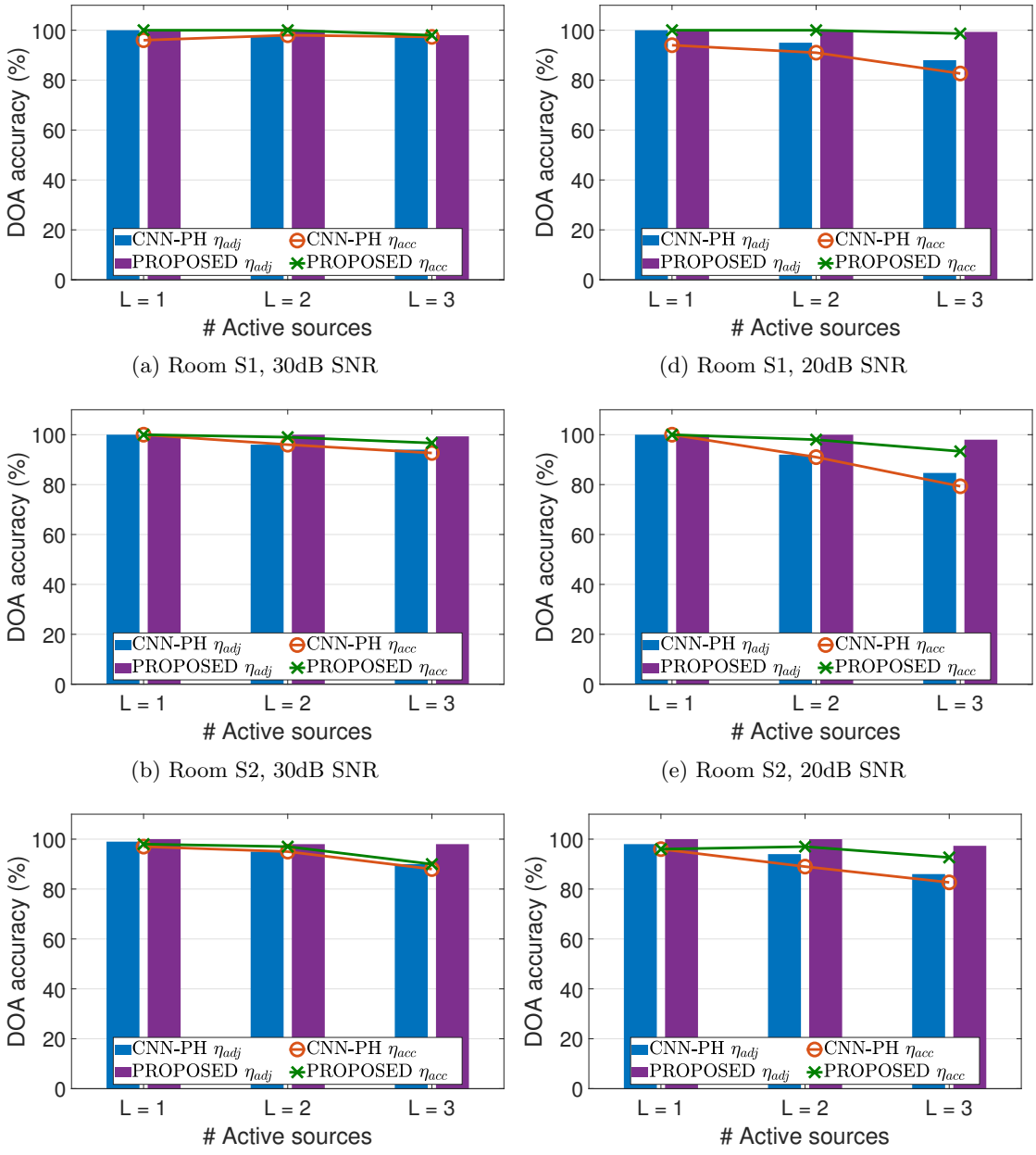
Azimuth estimation for a fixed elevation

In the first set of experiments, we considered uniformly spaced azimuth points at 10° interval (i.e., $J = 36$) on the fixed elevation plane at 45° for both training and testing. Hence, we can use (6.24) to determine the angular separation between the adjacent classes as $\Delta_\Omega = 7.07°$. For each room, we emulated 2 different signal to noise ratio (SNR) by adding white Gaussian noise and evaluated the performance for up to 3 active sources, i.e., $L = [1, 3]$. As "CNN-PH" was originally designed to be trained for all possible angular combinations in an L-source DOA estimation, we trained "CNN-PH" for 36 and 1260 unique angular combinations based on 36 azimuth classes for $L = 1$ and 2, respectively[4]. However, for testing with $L = 3$, we trained "CNN-PH" for 2-source mixture (1260 angular combinations) to understand the performance in a dynamic acoustic scenario. In contrast, the proposed method was always trained for the single-source scenario, e.g., 36 unique cases for this experiment, irrespective of the number of sources in the testing environment.

Fig. 6.3 shows DOA accuracy of the competing methods under different scenarios. At SNR = 30dB in Fig. 6.3, we observe that both the methods perform well for $L = 1$ and 2 although the proposed method consistently exhibits slightly

[3]As an instance, the average size of the training dataset in Section 6.4.3 was $10,505$ samples per label, each being a $[4 \times 4]$ matrix.

[4]The training process for "CNN-PH" is outlined in [150, pp. 13]



(a) Room S1, 30dB SNR

(d) Room S1, 20dB SNR

(b) Room S2, 30dB SNR

(e) Room S2, 20dB SNR

(c) Room S3, 30dB SNR

(f) Room S3, 20dB SNR

Figure 6.3: Azimuth estimation under different simulated reverberant and noisy environments on a 45° elevation plane.

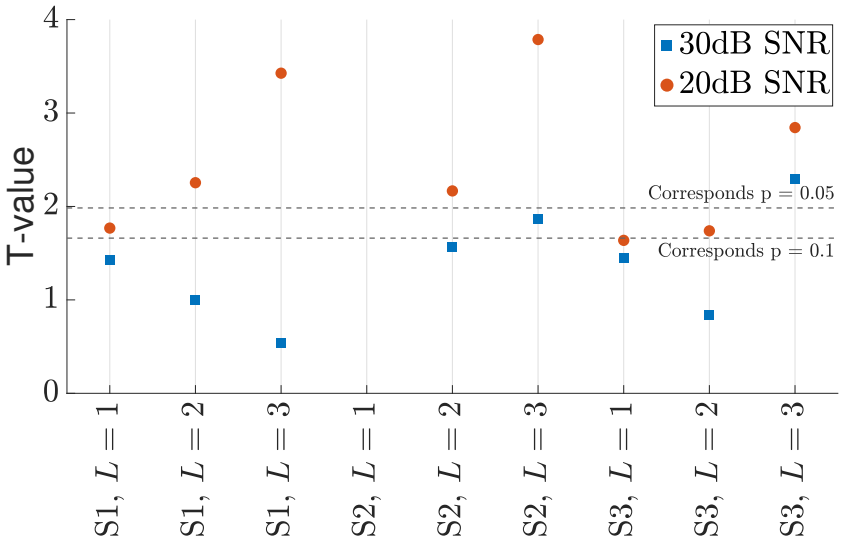
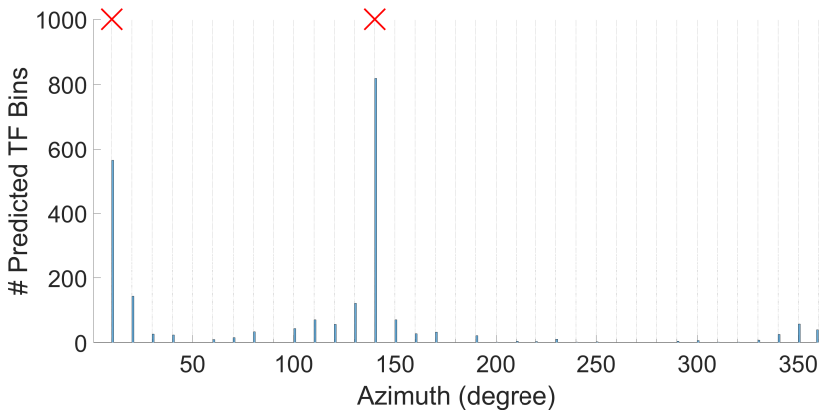


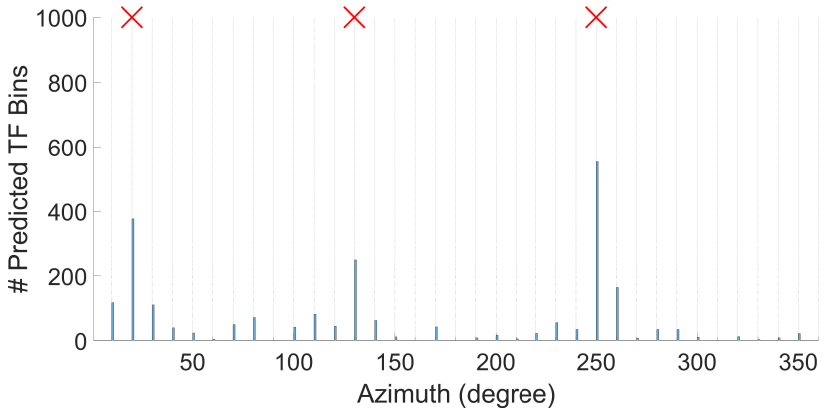
Figure 6.4: T-values in each of the scenarios of Fig. 6.3 calculated based on two-tailed independent samples t-tests. T-values above the reference dotted lines implies statistical significance based on corresponding p-values.

better performance. For 3-source combination under stronger reverberation, the proposed method holds the level of adjacent accuracy η_{adj} while "CNN-PH" shows performance degradation for both the metrics. For the noisy environments of SNR = 20dB in Fig. 6.3, the performance distinctions are more prominent where the proposed algorithm outperforms "CNN-PH" in each scenario. The use of modal coherence as learning feature ensures steady performance of the proposed algorithm at low SNR. We can also observe "CNN-PH" suffers significant performance issues for $L = 3$ due to the fact that we did not train "CNN-PH" for all possible 3-source combinations. We also investigated statistical significance of the results shown in Fig. 6.3 through two-tailed independent samples t-tests. The null hypothesis for the t-test was that both "CNN-PH" and the proposed algorithm have the equal mean estimation error. The results of the t-tests, shown in Fig. 6.4, rejects the null hypothesis for the scenarios where Fig. 6.3 exhibits notable difference between the competing methods. This implies that the results we present in Fig. 6.3 are statistically significant and can be considered as the usual outcome.

For reference, Fig. 6.5 plots the TF bin prediction histogram in room S3 for $L = 2$ and 3 along with the true azimuths. The histogram shows a clear peak



(a) $L = 2$



(b) $L = 3$

Figure 6.5: TF bin prediction histogram in room S3 ($T_{60} = 500$ms). Red crosses denote the ground truths.

at each true azimuth location which can be separated using a suitable clustering algorithm.

Performance in a practical room with babble noise

We conducted the next set of experiments in a big hall with strong reverberation, we named it room P1 in Table 6.2. The room schematic for P1 is shown in Fig. 6.6. The recording was performed with an *Eigenmike* [199], however, only first-order harmonics were used for this task. The source was placed at a 2.8m distance from the array in a uniform azimuth grid of 30° interval ($J = 12$) on a 95° elevation plane. Directional babble noise was added to the recordings at 10dB SNR from

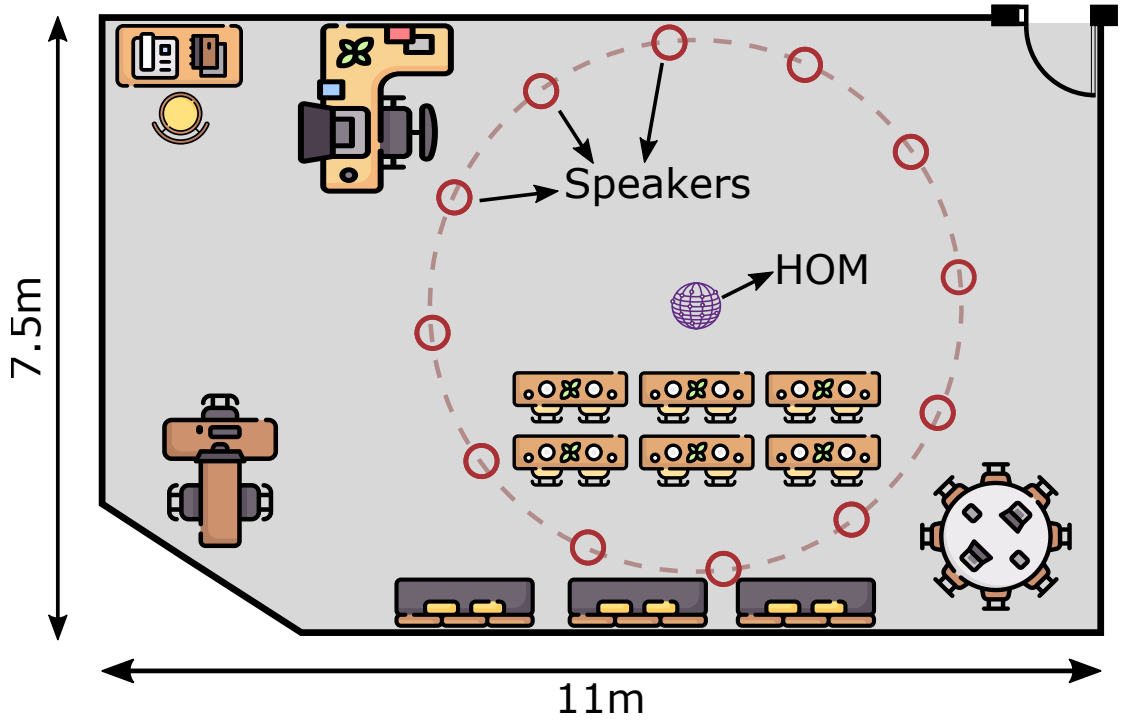


Figure 6.6: An artist's impression of room orientation and experimental setup for the practical room P1 (Table 6.2). The sound sources were placed 2.8 m from the microphone array.

multiple random locations. This time we trained "CNN-PH" with all possible angular combinations for both $L = 2$ (132 angular combinations) and $L = 3$ (1320 angular combinations) while the proposed method used the same strategy of single-source training for 12 classes. The comparative performance is shown in Fig. 6.7 where the proposed algorithm shows a significantly better accuracy than "CNN-PH", especially for $L = 3$, despite "CNN-PH" being trained for all possible angular combinations in each case. Note that, we found no significant performance improvement for higher harmonic orders. This can be due to the low spatial resolution of the training data. The higher order modes can be useful with a denser source distribution, at high frequencies, or when a regression-based model is used.

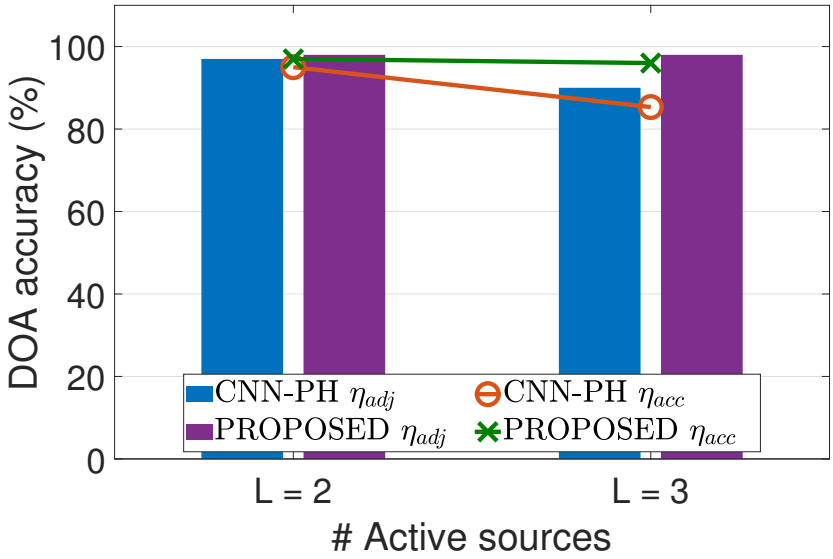


Figure 6.7: Azimuth estimation accuracy with practical recordings with babble noise at 10dB SNR. The tests were performed on 95° elevation plane.

Joint estimation of azimuth and elevation

So far, we have shown results only for azimuth estimation based on the proposition that the proposed algorithm can estimate azimuth and elevation simultaneously without interfering with each other. In this section, we are going to validate this proposition by performing a full DOA estimation in room S2. We designed a 3D uniform spatial grid with 30° resolution for azimuths ($J = 12$) and 20° resolution for elevations. Furthermore, we considered the elevation range $30° - 150°$. That makes a total of 7 unique elevation classes ($I = 7$) and a total 84 points on the 3D DOA grid. The rest of the simulation criteria remain the same as Section 6.4.3.

We slightly modified the CNN architecture for this section to accommodate the joint estimation in an efficient manner. As in the previous experiments, we calculated the feature snapshot for each TF bin, but this time we labelled them separately for azimuth and elevation. The CNN architecture remains the same for the most part except at the last layer when we branched out 2 identical but separated fully connected heads and supplied them with azimuth and elevation labels, respectively. Hence, at the testing stage, the system outputs two separate prediction sets for azimuth and elevation - one from each separated head. Note that, due

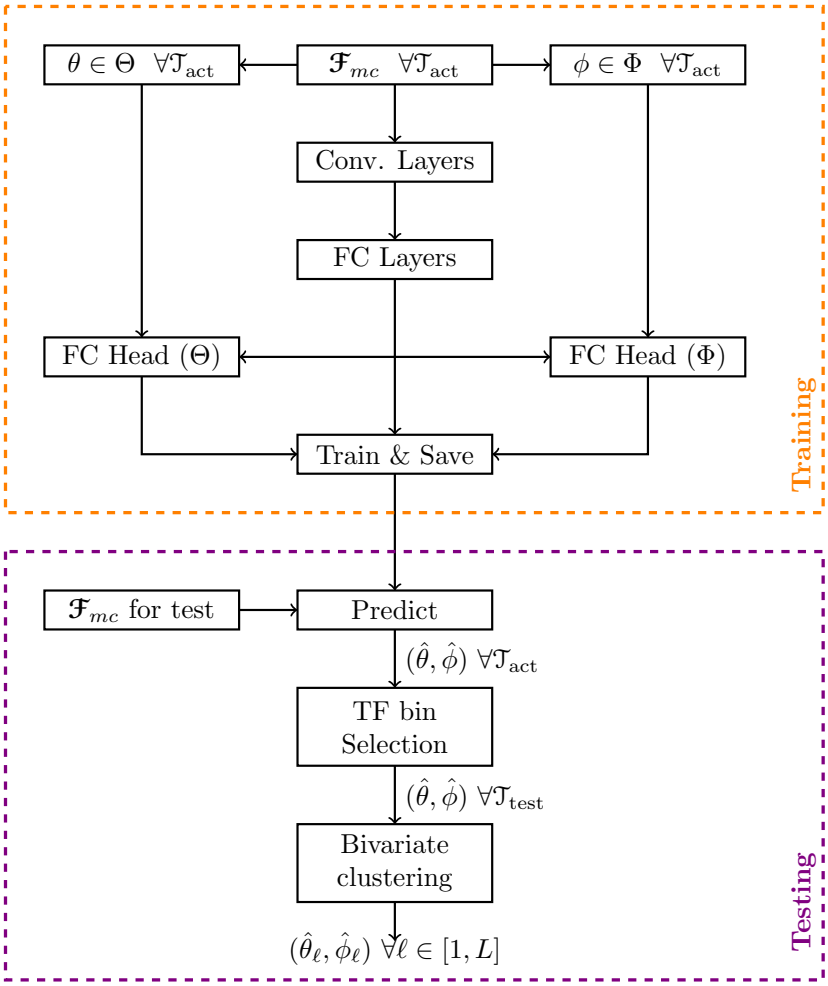
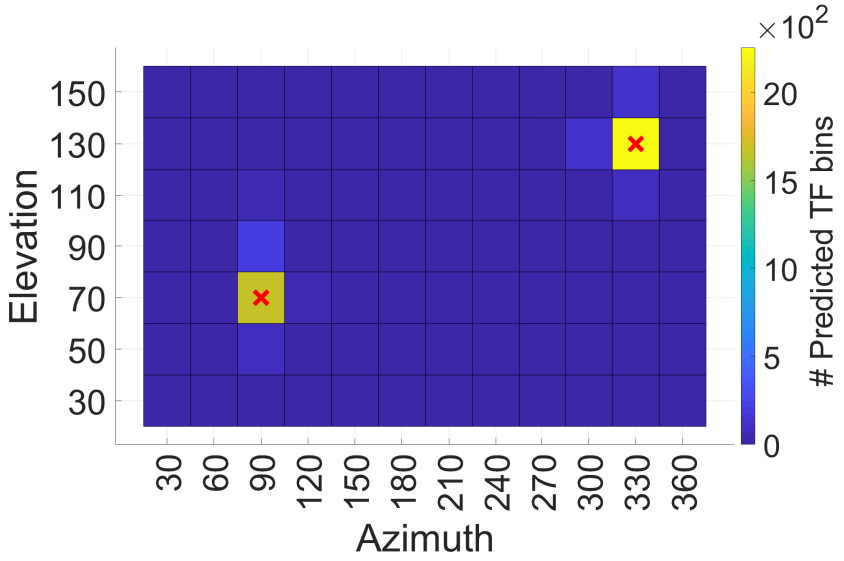


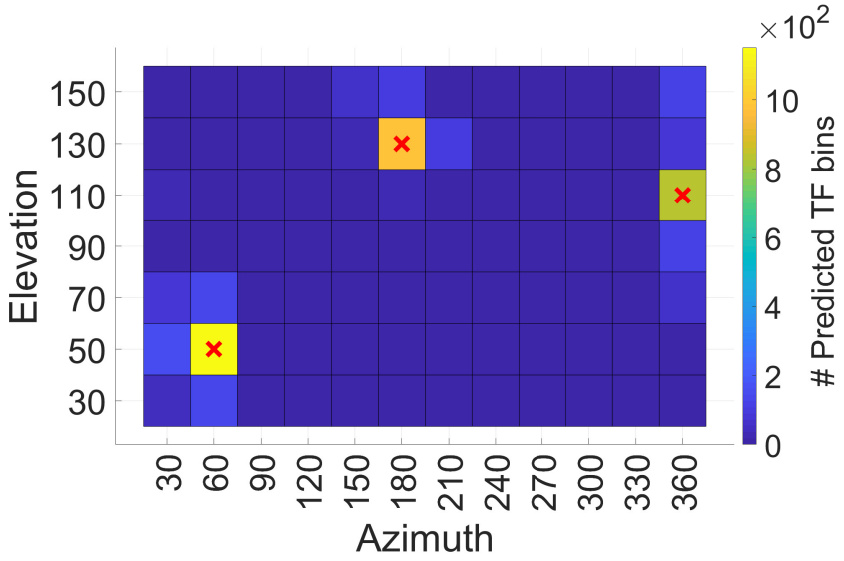
Figure 6.8: A block diagram for joint estimation of azimuth and elevation.

to the independent estimation strategy, it is important to jointly pack the predicted azimuths and elevations for each TF bin so that the estimated angles from the same source remain together. Subsequently, we clustered the prediction outcomes using the bi-variate k-means clustering algorithm. A block diagram for the joint estimation of azimuth and elevation is shown in Fig. 6.8[5]. It is worth mentioning that the proposed algorithm can readily be expanded for full source localisation through additional training for radius-dependency or corresponding Cartesian coordinates due to its ability of independent estimation of different location parameters.

[5]Fig. 6.8 does not necessarily depict the actual processing flow, rather a visual aid for understanding the task.



(a) $L = 2$



(b) $L = 3$

Figure 6.9: Color map for joint estimation of azimuth and elevation with the proposed method in room S2 at 30dB SNR. Red crosses denote the ground truths. The accuracy of a joint estimation over 50 tests was found to be similar compared to the standalone azimuth estimation.

The outcome of the experiments is shown in Fig. 6.9 in terms of coloured heat map based on the number of predicted TF bins for each DOA. It is clear from the

figure that the proposed method had no difficulties in predicting full DOA in the same manner as with the azimuth predictions. More importantly, the accuracy of the joint DOA estimation was found to be similar to that of a standalone azimuth estimation of Fig. 6.3 (and hence, the accuracy plots are not shown separately for this section). This is an expected behaviour, as the azimuth and elevation estimation processes are independent and should not be affected due to the joint processing.

Azimuth estimation on a different elevation plane

In this section, we analyse the performance of the proposed algorithm when training and testing were performed on different elevation planes. For the purpose of this section, we used room S2 at 30dB SNR. The tests were performed for sources on 60° elevation plane while the training data were obtained from a different elevation. In Fig. 6.10, we show the estimation accuracy for 2 distinct cases - when training data were obtained from (case 1) 45° elevation plane only and (case 2) 45° and 75° elevation planes. We observe a clear improvement in case 2 over case 1 due to the fact that when we trained the network on 2 different elevation planes, the model learned the evolution of feature for change in elevation and predicted azimuths in an arbitrary elevation plane more accurately. As the machine learning algorithms take a data-driven approach, it is possible to further improve the performance by training on additional planes.

It is worth noting that the proposed feature snapshot $\boldsymbol{\mathcal{F}}_{\text{mc}}$ is mostly comprised of the spherical harmonics where the dependency on θ and ϕ come through independent Legendre and exponential functions, respectively, as shown in (2.13). Therefore, the impact of elevation change on $\boldsymbol{\mathcal{F}}_{\text{mc}}$ comes mainly as a constant scaling factor. For this reason, even for case 1 when training and testing were done in separate individual elevation planes, the model did not entirely fail, rather gets confused by the reverberation, noise and other non-linear distortions. This is apparent from Fig. 6.10 where we observe a better accuracy in terms of η_{adj} but a significant difference with η_{acc}.

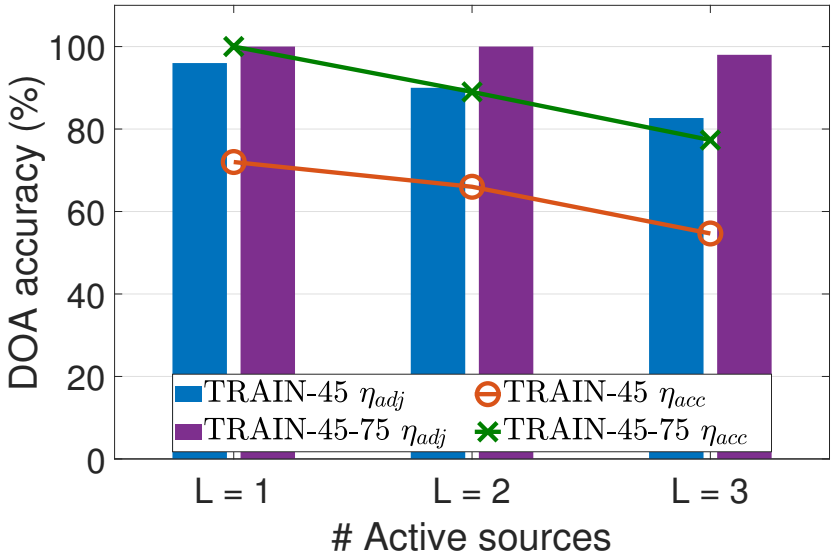


Figure 6.10: Azimuth estimation on 60° elevation plane when training was performed on different elevation plane. TRAIN-45 denotes the case when training was performed on 45° elevation only whereas for TRAIN-45-75, training was performed with data from 45° and 75° elevations.

Impact of source to microphone distance

We investigate the impact of the varying source to microphone distance on the proposed algorithm. We used the same simulated room S2 with the exception that we increased the dimension of the room to $[8 \times 8 \times 4]$m for this particular section in order to have a larger range for distance variation. The microphone array position remained at the centre of the room, however, we varied the source position between $0.5-3$ m from the microphone array. The training was performed at a fixed distance of 1m. The plots in Fig. 6.11 suggests that there is no significant change in estimation accuracy for varying source to microphone distances during the test. This implies that our model doesn't need to be trained separately for near-field and far-field considerations due to the separation of radius-dependency from DOA-dependency in the modal coherence model. To understand the behaviour, we examine the analytical expression of α_{nm} for the direct path for ℓ^{th} source [185, pp. 31]

$$\alpha_{nm}^{(\ell)}(k,r) = ikh_n(kr_\ell)Y_{nm}^*(\widehat{\boldsymbol{x}}_\ell) \tag{6.27}$$

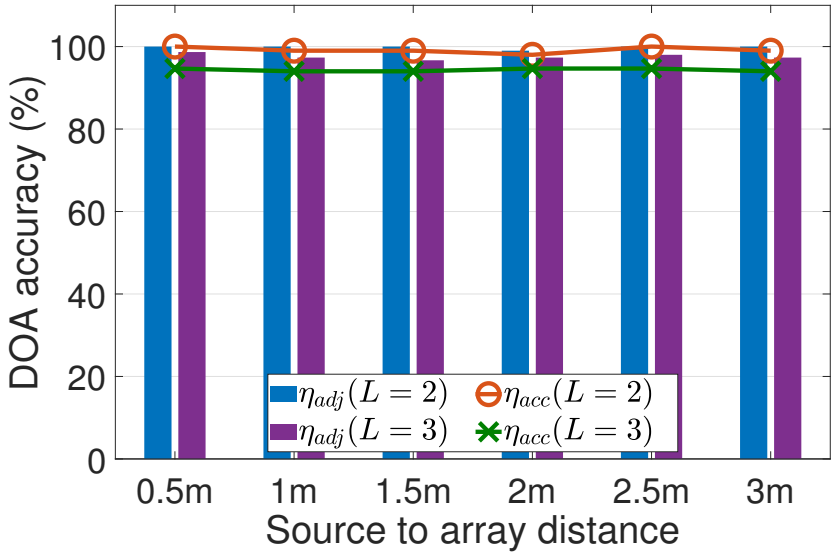


Figure 6.11: Azimuth estimation performance of the proposed algorithm with different source to microphone distances. The training was performed with sources at 1m distance from the array centre on a 45° elevation plane.

where $h_n(\cdot)$ is the spherical Hankel function of the first kind. From (6.27) it is clear that the radial dependency comes through the Hankel function $h_n(kr_\ell)$ together with the frequency-dependent k. As we trained our model for different frequencies, the impact of varying $h_n(kr_\ell)$ on the feature pattern is already captured during the training even with a fixed radius, hence, any radial change does not pose a major threat to the performance of the proposed algorithm.

Number of active sources

In the last set of experiments, we tried increasing the number of sources on the same 45° elevation plane in the acoustic scene of room S2 at 30dB SNR. As we observe in Fig. 6.12, the accuracy gradually decreases with the increasing number of sources. The performance issue can be contributed by an increased violation of W-disjoint orthogonality with an increased number of sources. However, examining the histograms of random individual tests, we also found many instances when the performance degradation was caused by the failure of the k-means clustering algorithm and the ambiguity in the histogram for nearby sources. It is possible

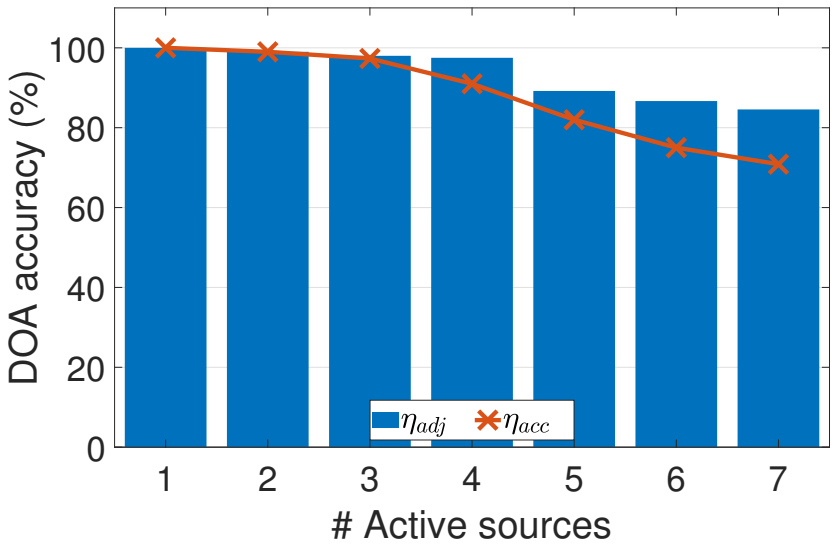


Figure 6.12: Azimuth estimation performance of the proposed algorithm with different number of active sources on a 45° elevation plane.

to improve the performance with a careful selection of a more robust clustering algorithm, however, the investigation for a better clustering algorithm is out of the scope of this work. To avoid ambiguity due to nearby sources, we can impose a restriction for maintaining a minimum distance between two sources before applying this algorithm.

6.5 Summary

In this chapter, we proposed a modal coherence-based feature to train a convolutional neural network for DOA estimation realised in the spherical harmonic domain. The chapter offers the following major contribution:

- We proposed a CNN-based framework to estimate DOAs of simultaneously active multiple sound sources. We used a novel feature to train a CNN which utilises the modal coherence of a reverberant sound field in the spherical harmonic domain. We showed that modal coherence represents a unique pattern for each direction which can be learned and used for estimating source DOAs in a composite acoustic scene.

- We offered a new perspective by introducing a single-source training scheme for multi-source localisation in a reverberant environment. The proposed strategy saves significant time and resources during the training stage as well as allows us to reuse the same trained model during the testing stage irrespective of the number sources in the acoustic mixture.
- Unlike the existing methods in the CNN domain, we treat individual STFT bin separately for training. Hence, the proposed method is capable of performing parallel azimuth and elevation estimation, hence, allows us to perform full DOA estimation without affecting the estimation accuracy compared to standalone azimuth estimation.
- The data-driven approach of DOA estimation overcomes some of the fundamental limitations of the traditional parametric domain approaches.
- The proposed algorithm outperforms the competing state-of-the-art technique while having a significantly reduced resource requirements.

In summary, we contributed to the cause by proposing a method that performs better than the contemporary CNN-based methods in various acoustic environments, requires significantly less resource and time for training, and predict the DOA based on a single training model for a room irrespective of the number of sources.

6.6 Related Publication

- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "Multi-Source DOA Estimation through Pattern Recognition of the Modal Coherence of a Reverberant Soundfield", *IEEE/ACM Transactions on Audio Speech and Language Processing*, Volume 28, pp. 605–618, 2019.

This page intentionally left blank.

Chapter 7

Post-filter Selection for Non-Orthogonal Signals

Choosing an appropriate post-filter is an important consideration in the proposed source separation techniques, as outlined in Fig. 1.4, to improve the perceived quality. In Chapter 5, we employed a traditional Wiener filter to enhance source separation performance (Fig. 5.1 and 5.9). In this chapter, we review various existing options for spectral enhancements in addition to the Wiener Filter and explore their strengths and limitations. Most of the existing post-filtering techniques use the strict assumption of orthogonality between the desired and undesired signal components. However, in a practical scenario, the orthogonality assumption is often violated due to a short processing window as well as the correlated reflected signals from the surrounding objects. The scale and nature of the error due to this non-orthogonality depend on the acoustic condition and the underlying constraints of the techniques. Hence, it is important to analyse the performance of the existing post-filtering solutions under different reverberant and noisy environments to be able to make an informed decision while choosing a post-filtering method for source separation in a reverberant environment.

7.1 Introduction

Spectral enhancement improves the interference rejection of a mixed signal and is typically achieved by designing a suitable post-filter in the frequency domain. Several post-filtering techniques are available for this purpose such as a Wiener filter (WF) [213], spectral subtraction (SS) [103], short-time spectral amplitude estimator (STSA) [104], and log-spectral amplitude estimator (LSA) [105]. One common assumption made by all of these conventional approaches is the strict orthogonality between the desired and undesired components of an observation. Based on this orthogonality assumption, each of these methods chose different approaches to devise the corresponding transfer function for a spectral filter. However, this orthogonality assumption does not hold in many practical scenarios due to limited time-domain support and the short-time stationarity of the speech signals. Hence, the performances of these post-filtering methods vary under different acoustic environments depending on the robustness of their underlying constraints in that specific scenario. To improve the performance of a conventional SS-based post-filter, Lu *et al.* relaxed the orthogonality assumption by proposing a geometric approach to spectral subtraction (GSS) [153], however, that introduces additional constraints in their formulation. Originally [153] was assessed against background noise suppression, however, the deviation from orthogonality assumption becomes more prominent for reverberant environments due to correlated reflections from the surrounding objects. In this chapter, we review the conventional techniques and their performance and analyse the GSS model by scrutinising its transfer function to gain an insight into the algorithm.

In order to conduct a comparative analysis between the conventional and contemporary approaches using a standard methodology, we evaluate the performance of the post-filters in terms of single-channel dereverberation and noise suppression using the REVERB challenge 2014 dataset [214]. The REVERB challenge 2014 dataset does not offer a suitable option for spherical harmonic decomposition, hence, we cannot apply the power spectral density (PSD) estimation technique we developed in Chapter 4. Instead, we opt for a traditional approach to estimate noise and reverberation PSDs and design the evaluation system based on that. As we intend to evaluate the performance of the post-filtering techniques, the lack of

spatial domain processing does not hinder our objective. Note that, though this analysis is based on single-source dereverberation and noise suppression, the insight we gain is equally applicable for a source separation scenario because of the existence of reverberation and background noise. Furthermore, understanding of the performance limitations due to non-orthogonality in a short-time Fourier transform (STFT) frame is useful irrespective of whether the undesired signal is coming from a separate source or from its own reflections. At the end of this chapter, we hope to acquire the knowledge about the relative advantages between different spectral filters and their limitations, which help us to make an informed decision on selecting an appropriate post-filtering technique.

7.2 Problem Statement

The input to the Wiener filter in Fig. 5.1 and 5.9 was a single-channel signal from the beamformer output and the centre-microphone, respectively. Hence, in this chapter, we consider a single-channel audio $y(n)$ as our input signal while analysing the performance of different post-filtering techniques. Furthermore, in the context of Chapter 5, $y(n)$ constituted of the desired and undesired speech, reverberation, and noise signals. However, to compare the performance of different post-filters with the contemporary methods of the REVERB challenge 2014, we consider single-source noisy and reverberant environment provided by the REVERB challenge 2014 dataset. Note that, as our objective is to measure the performance variation of different post-filters in terms of interference rejection, the discussion and findings of this chapter is equally applicable for the post-filter selection strategy of Chapter 5. The remaining of this chapter focuses on $y(n)$ as the signal of interest and study the performance of different signal enhancement techniques.

Under the single-source assumption, the input signal $y(n)$ to the post-filter is given by

$$y(n) = x_e(n) + x_r(n) + x_n(n) \tag{7.1}$$

where $x_e(n)$ is the desired signal, $x_r(n)$ represents the undesired reverberation component, and $x_n(n)$ is the background noise. Our goal is to compare the performances of different post-filtering techniques in extracting the desired signal $x_e(n)$

from $y(n)$. The processing is done in the STFT domain where (X_e, X_r, X_n) represents the corresponding STFT coefficients of (x_e, x_r, x_n).

7.3 Spectral Enhancement

In this section, we briefly describe the transfer functions of different post-filtering techniques that we intend to compare.

7.3.1 The conventional approaches

For the theoretical analysis, we consider the following simplified signal model in frequency domain

$$P_m(\omega) = P_d(\omega) + P_i(\omega) \tag{7.2}$$

where the desired signal $P_d(\omega)$ is distorted by the interfering signal $P_i(\omega)$ to produce measured signal $P_m(\omega)$ with ω being the angular frequency. In the context of Section 7.2, $P_d(\omega) = X_e(\omega)$ whereas $P_i(\omega)$ can be in the form of the background noise $X_n(\omega)$ or the late reverberation $X_r(\omega)$ modelled as an additive disturbance. In the subsequent sections, we discuss two widely-used conventional post-filtering techniques in terms of their transfer functions. For the remainder of this section, we frequently omit ω in the definitions for brevity.

Conventional Spectral Subtraction

The conventional SS is based on the instantaneous signal or power spectra. From (7.2), the squared-magnitude spectrum of measured P_m is given by

$$|P_m|^2 = |P_d|^2 + |P_i|^2 + 2\ |P_d|\ |P_i|\ \cos\left(\theta_{di}\right) \tag{7.3}$$

where θ_{di} is the phase difference between P_d and P_i, and $|\cdot|$ denotes the absolute value. Spectral subtraction assumes P_d and P_i to be orthogonal, i.e. $\theta_{di} = \pi/2\ \forall\omega$. Hence, the cross-terms of (7.3) become zero and the transfer function of SS is given by

$$H_{ss}(\omega) = \sqrt{1 - \frac{1}{\gamma(\omega)}} \tag{7.4}$$

where $\gamma(\omega) = \frac{|P_m(\omega)|^2}{|P_i(\omega)|^2}$ is the *a posteriori* signal to interference ratio (SIR) based on the instantaneous signal spectra. Note that, $H_{ss}(\omega)$ is a real quantity, hence, the noisy phase remains unchanged in signal reconstruction[1].

Wiener Filter

In the Wiener filter theory, an optimum filter H_w is designed to estimate P_d by minimising the mean squared error of the estimation, and the solution takes the form of [216]

$$H_w(\omega) = \frac{\Phi_{dm}(\omega)}{\Phi_m(\omega)} \tag{7.5}$$

where Φ_{dm} is the cross spectral density (CSD) between P_d and P_m whereas Φ_m is the PSD of P_m. Φ_m and Φ_{dm} can be calculated from (7.2) as

$$\Phi_m = \Phi_d + \Phi_i + \Phi_{di} + \Phi_{id} \tag{7.6}$$

$$\Phi_{dm} = \Phi_d + \Phi_{di} \tag{7.7}$$

where Φ_d and Φ_i are the PSDs of P_d and P_i, respectively, and Φ_{di} and Phi_{id} are the CSDs between P_d and P_i. In the Wiener solution, the cross-terms are considered zero and hence, using (7.5), (7.6) and (7.7), the solution becomes

$$\begin{aligned} H_w(\omega) &= \frac{\Phi_d(\omega)}{\Phi_d(\omega) + \Phi_i(\omega)} \\ &= \frac{\widetilde{\xi}(\omega)}{\widetilde{\xi}(\omega) + 1} \end{aligned} \tag{7.8}$$

where $\widetilde{\xi}(\omega) = \frac{\Phi_d(\omega)}{\Phi_i(\omega)}$ is the *a priori* SIR.

7.3.2 Limitations of the conventional approaches

The assumption of uncorrelated desired and undesired signal implies that

$$\mathbb{E}\big\{P_d(\omega)P_i^*(\omega)\big\} = 0 \tag{7.9}$$

[1]It has been argued in literature that the phase distortion is largely inaudible in audio processing [215]

Table 7.1: Cross-term estimation error (ϵ_{avg}) in dB under different reverberation time (T_{60}), noise type, SIR, and window length of STFT.

	Frame Size			
	8 ms	16 ms	32 ms	64 ms
Reverberant speech				
$T_{60} = 300$ ms	-8.37	-8.5	-8.78	-8.96
$T_{60} = 600$ ms	-2.9	-2.88	-2.81	-3.22
$T_{60} = 700$ ms	-1.82	-1.77	-1.42	-2.44
Speech with air-condition noise				
SIR = 10 dB	-4.66	-4.85	-4.69	-3.98
SIR = 0 dB	0.34	0.15	0.31	1.02
Speech with white noise				
SIR = 10 dB	-4.93	-4.81	-4.07	-3.65
SIR = 0 dB	0.1	0.21	0.95	1.21

where $\mathbb{E}\{\cdot\}$ denotes the expected value and $(\cdot)^*$ represents complex conjugation. However, (7.9) does not hold for late reverberation component of a signal which exhibits finite correlation with the source. Furthermore, in practical experiments, (7.9) needs to be approximated from the finite window length to avoid smearing effect of non-stationary speech signal. Such a truncation results in residual cross-terms of notable magnitude that leads to overestimation or underestimation of the estimated signal inside the short processing frame.

Lu and Loizou used a noisy environment to demonstrate that the cross-term error due to non-orthogonality is particularly large around 0 dB *a priori* SIR and more severe as θ_{di} approaches π [153]. To study the cross-term error in different noisy and reverberant environments, we define the following error definition

$$\epsilon_{\text{avg}} = \frac{1}{T}\sum_{\forall\tau}\frac{\sum\limits_{\forall k}\left||P_d(\tau,k)|^2 - |\widehat{P}_d(\tau,k)|^2\right|}{\sum\limits_{\forall k}|P_d(\tau,k)|^2} \tag{7.10}$$

$$= \frac{2}{T}\sum_{\forall\tau}\frac{\sum\limits_{\forall k}|P_d(\tau,k)|\ |P_i(\tau,k)|\ \left|\cos\left(\theta_{di}(\tau,k)\right)\right|}{\sum\limits_{\forall k}|P_d(\tau,k)|^2} \tag{7.11}$$

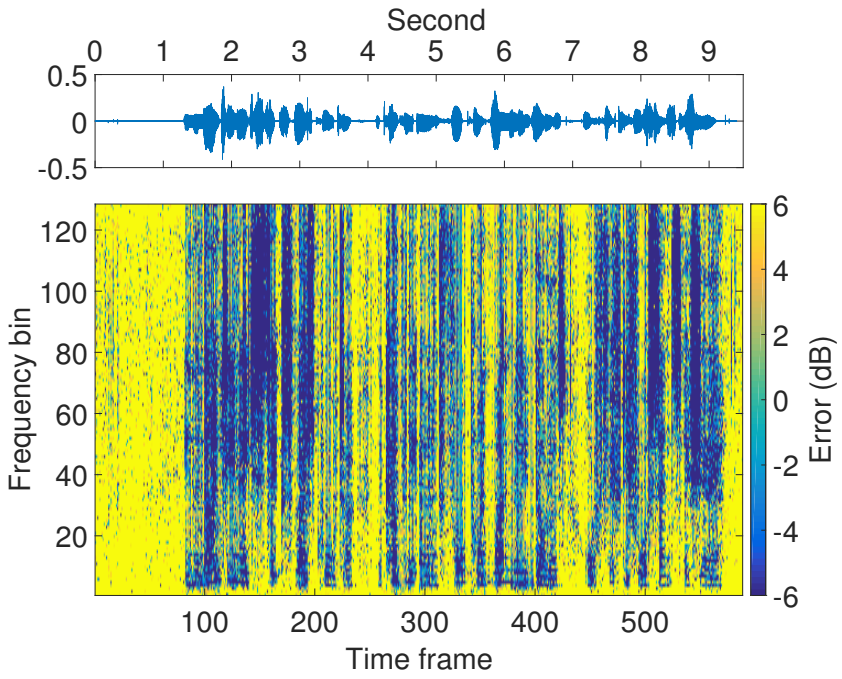


Figure 7.1: Cross-term error in a speech signal with 16 ms frames and no overlap, at 10 dB *a priori* SIR

where τ and k are the time and frequency bins of STFT, respectively, and T is the total number of time frames. Note that, Eq. (7.11) is obtained from (7.10) using (7.3) and (7.4). The results shown in Table 7.1 were obtained from the voiced-frames of 25 random speech signals assuming oracle PSD knowledge. The exclusion of the unvoiced frames was done to avoid error magnification in the frames where $X_e(\tau, k)$ is very small.

The results of Table 7.1 confirms the presence of significant cross-term error within a short STFT frame, especially with strong reverberation or background noise. Furthermore, a large cross-term error with both reverberant and noisy signal suggests that the non-zero cross-terms are majorly contributed from the lack of orthogonality in the short STFT frame. The spectrum of the cross-term estimation error is shown in Fig. 7.1 for reference.

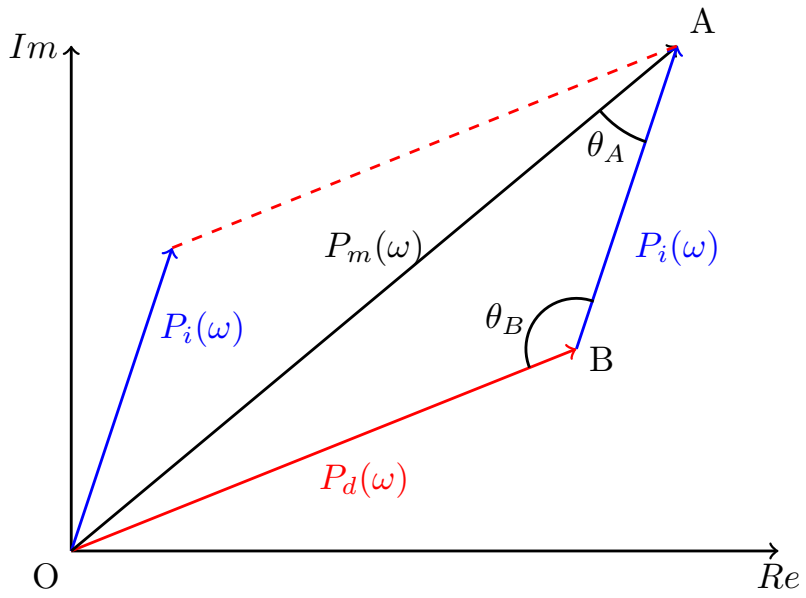


Figure 7.2: Phasor diagram of (7.2)

7.3.3 Geometric spectral subtraction

To circumvent the cross-term estimation error, an alternative approach to SS was proposed in [153] based on vector geometry. The authors used the phasor diagram of the signal model in (7.2), as shown in Fig. 7.2, to devise a GSS transfer function utilising certain trigonometric identities. Using the law of Sines in the $\triangle OAB$ of Fig. 7.2, we get

$$\frac{|P_d(\omega)|}{|P_m(\omega)|} = \frac{\sin\left(\theta_A(\omega)\right)}{\sin\left(\theta_B(\omega)\right)}. \tag{7.12}$$

Also the laws of Cosines in $\triangle OAB$ of Fig. 7.2 lead to

$$\cos\left(\theta_A(\omega)\right) = \frac{|P_m(\omega)|^2 + |P_i(\omega)|^2 - |P_d(\omega)|^2}{2\,|P_m(\omega)|\;\;|P_i(\omega)|} \tag{7.13}$$

$$\cos\left(\theta_B(\omega)\right) = \frac{|P_d(\omega)|^2 + |P_i(\omega)|^2 - |P_m(\omega)|^2}{2\,|P_d(\omega)|\;\;|P_i(\omega)|}. \tag{7.14}$$

Combining (7.12), (7.13), and (7.14), the gain function for GSS is formulated as [153]

$$H_{gs} = \frac{|P_d(\omega)|}{|P_m(\omega)|} = \sqrt{\frac{1 - \frac{[\gamma(\omega)+1-\xi(\omega)]^2}{4\gamma(\omega)}}{1 - \frac{[\gamma(\omega)-1-\xi(\omega)]^2}{4\xi(\omega)}}} \tag{7.15}$$

where $\xi(\omega) = \frac{|P_d(\omega)|^2}{|P_i(\omega)|^2}$ is the *a priori* SIR based on instantaneous spectra.

7.3.4 Limitation of geometric spectral subtraction

Eq. (7.15) requires both *a priori* and *a posteriori* SIRs, similar to [104], [105], however, differs in the sense that, it was formulated with the instantaneous signal spectra like the conventional SS. Under the assumption of the zero cross-terms, we can draw the following relationship between $\xi(\omega)$ and $\gamma(\omega)$ from (7.3):

$$\xi(\omega) = \gamma(\omega) - 1, \tag{7.16}$$

in which case the GSS becomes identical to the conventional SS. Note that, GSS transfer function can be simplified as

$$H_{gs} = \sqrt{\frac{\xi(\omega)}{\gamma(\omega)} \frac{4\gamma(\omega) - [\gamma(\omega) + 1 - \xi(\omega)]^2}{4\xi(\omega) - [\gamma(\omega) - 1 - \xi(\omega)]^2}} \tag{7.17}$$

$$= \sqrt{\frac{\xi(\omega)}{\gamma(\omega)}}, \text{ for } \xi(\omega) \neq \left[\sqrt{\gamma(\omega)} \pm 1\right]^2. \tag{7.18}$$

While SS depends on the *a posteriori* SIR, GSS performance is limited by both ξ and γ. Hence, GSS is more sensitive to PSD estimation error compared to SS. ξ is commonly estimated with a decision-directed approach [104] as a weighted average of the past value and present *a posteriori* SIR. Therefore, the performance of GSS is largely affected by the choice of the weighting factor, as we will show in the results section. Hence, it is fair to conclude that while GSS offers an improvement on estimation accuracy by incorporating the cross-terms, its performance can deteriorate quickly in the presence of a significant PSD estimation error. Otherwise, it is expected that in scenarios that involve high cross-term components, GSS should outperform conventional SS. We shall validate this statement in the results section.

A direct theoretical comparison between the GSS and WF is difficult to make as they are based on different underlying principles. Hence, we depend on the

simulation results to compare the performances of GSS and WF.

7.4 System Model

In Chapter 5, we utilised Wiener filter as a post-filter where the estimated PSD components from the multi-channel PSD estimation block served as the input (Fig. 5.1 and 5.9). Our main objective in this chapter is to study different post-filtering solutions and analyse their performance independent of the underlying PSD estimation technique. We use the single-channel speech database of the REVERB challenge 2014 [214] in order to make a comparative analysis between the conventional and contemporary spectral filters. To that end, we rely on simple single-channel techniques to estimate desired and undesired PSD components from a mixed recording. Subsequently, we use the estimated PSDs to calculate *a priori* and *a posteriori* SIRs, γ, ξ, and $\widetilde{\xi}$, and corresponding transfer functions for each of the analysing post-filters based on the formulation outlined in Section 7.3. Finally, the noisy and reverberant speech signals from the REVERB challenge 2014 database is processed separately using each of the post-filters to extract the clean speech and the outputs of the filters are compared using industry-standard metrics.

7.4.1 Single-channel PSD estimation

We use Lebart's late reverberation energy estimator [109] based on Polack's statistical model of RIR [217] to estimate reverberant PSD as

$$\Phi_r(\tau, k) = e^{-2\Delta N_e}\ \Phi_{er}(\tau - \frac{N_e}{N_{\text{hop}}}, k) \tag{7.19}$$

where Φ_{er} is the total signal PSD excluding noise, N_e is the length of early reflections, N_{hop} denotes STFT hop size, and

$$\Delta = \frac{3\ \log_e(10)}{T_{60}\ f_s} \tag{7.20}$$

with f_s being the sampling frequency. The reverberation time T_{60} is calculated using a maximum likelihood based estimator described in [218]. As Φ_{er} is not

available in the beginning, it is estimated from the measured signal using a recursive formula:

$$\Phi_{er}(\tau, k) \approx \eta_x \; \Phi_y(\tau - 1, k) + (1 - \eta_x) \; |Y(\tau, k)|^2 \tag{7.21}$$

where Φ_y is the PSD of the measured signal Y and $\eta_x \in [0, 1]$ is a smoothing factor.

Conversely, noise power is estimated from the speech pauses using the following moving average formula

$$\Phi_n(\tau, k) = \eta_v \; \Phi_n(\tau - 1, k) + (1 - \eta_v) \; |X_d(\tau, k)|^2 \tag{7.22}$$

where $X_d = X_e + X_n$ is the output of the dereverberation stage and $\eta_v \in [0, 1]$ is a smoothing factor with the constraint of $\eta_v = 1$ in the voiced frames.

7.4.2 *A priori* and *a posteriori* SIRs

SS and GSS require to compute the instantaneous *a priori* SIR ξ. However, due to the unavailability of the instantaneous spectra of desired and undesired signals, ξ is also estimated using a recursive formula. Furthermore, we observed that the recursion in the calculation of SIRs yields a better performance for both GSS and SS-based algorithms. This can be a result of a reduced error variance due to the smoothing operation. Hence, we use the following smoothed versions of the *a posteriori* SIRs

$$\gamma_1(\tau, k) = \beta \; \gamma_1(\tau - 1, k) + (1 - \beta) \; \min \left\{ \frac{|Y(\tau, k)|^2}{\Phi_r(\tau, k)}, \gamma_{max} \right\} \tag{7.23}$$

$$\gamma_2(\tau, k) = \beta \; \gamma_2(\tau - 1, k) + (1 - \beta) \; \min \left\{ \frac{|X_d(\tau, k)|^2}{\Phi_n(\tau, k)}, \gamma_{max} \right\} \tag{7.24}$$

where γ_1 and γ_2 are the *a posteriori* SIRs for dereverberation and noise suppression stage, respectively, $\beta \in [0, 1]$ is a smoothing constant, and γ_{max} is used to avoid over-attenuation of the signal.

The *a priori* SIRs ξ_1 and ξ_2 respectively for dereverberation and noise suppres-

sion stage are calculated based on a decision-directed approach [104], [153]

$$\xi_1(\tau,k) = \max\left\{\alpha\ \frac{|X_d(\tau-1,k)|^2}{\Phi_r(\tau-1,k)} + (1-\alpha)(\sqrt{\gamma_1(\tau,k)}-1)^2,\ \xi_{min}\right\} \tag{7.25}$$

$$\xi_2(\tau,k) = \max\left\{\alpha\ \frac{|\widehat{X}_e(\tau-1,k)|^2}{\Phi_n(\tau-1,k)|^2} + (1-\alpha)(\sqrt{\gamma_2(\tau,k)}-1)^2,\ \xi_{min}\right\} \tag{7.26}$$

where $\alpha \in [0,1]$ is a smoothing constant, $\widehat{X}_e$ is the estimated output of noise suppression block, and ξ_{min} represents the minimum allowed value for *a priori* SIR.

7.4.3 On two-stage approach of the solution

In the performance evaluation, we apply dereverberation task followed by a noise suppression block. A two-stage approach is used for a better voice detection and noise PSD estimation from the noisy measurements. As the late reverberation PSD overlaps with the signal PSD, the identification of the unvoiced frames can be erroneous and the PSD estimation accuracy may deteriorate. Hence, by suppressing the late reverberation component before performing the noise PSD estimation, we expect to increase the estimation accuracy.

However, there exists a contradictory argument regarding the sequence of the two-stage algorithm. The late reverberation energy estimator of (7.19) requires the noise-suppressed reverberant signal PSD Φ_{er} which remains unknown at the beginning. Hence, the late reverberation PSD needs to be estimated from the noisy PSD Φ_y which introduces additional error.

Hence, there has to be a compromise in determining the sequence of the dereverberation and noise suppression in a two-stage operation. Such a dilemma can be avoided with a joint PSD estimator such as the one we developed in Chapter 4.

7.5 Experimental Results

We measured the performance using the REVERB challenge 2014 speech enhancement (RCSE2014) dataset. We also used the official RCSE2014 evaluation tool for

Table 7.2: Room geometry and RT

Identifier	Source to microphone distance	RT
R1	50 cm	0.3s
R2	200 cm	0.3s
R3	50 cm	0.6s
R4	200 cm	0.6s
R5	50 cm	0.7s
R6	200 cm	0.7s

data generation and performance evaluation. RCSE2014 contains 6 different RIRs measured under different room conditions listed in Table 7.2. Note that, the data of Table 7.2 were not used in the experiments as per RCSE2014 guideline. The dataset contains 362 speech files with $f_s = 16$ kHz for each reverberant condition with a 20 dB background noise, unless mentioned otherwise.

7.5.1 Parameters settings & evaluation measures

We used a 16 ms Hanning window with 75% overlap for calculating a 256-point discrete Fourier transform. The parameter settings shown in Table 7.3 were chosen in an empirical manner based on a subset of the training data. We used perceptual evaluation of speech quality (PESQ) [204], speech to reverberation modulation energy ratio (SRMR) [219], frequency-weighted segmental SNR (FWSegSNR), cepstral distance (CD) and log-likelihood ratio (LLR) [193] for performance evaluation. For reference, higher PESQ, SRMR and FWSegSNR indicate a better performance, whereas the opposite is true for CD and LLR.

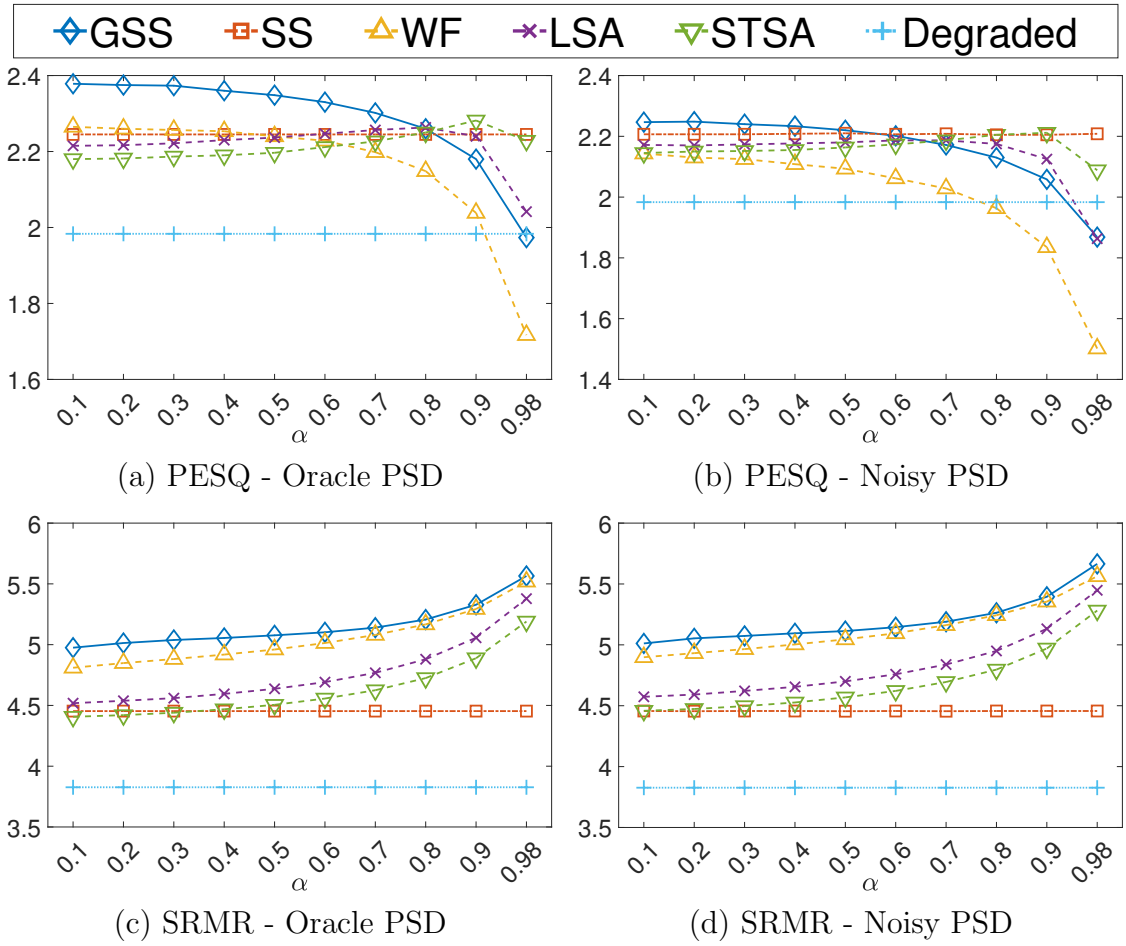
Table 7.3: Parameter settings used in the simulation

Parameter	**Value**
α	See Section 7.5.2
β, η_v	0.6
η_x	0.85
γ_{mx}	13 dB
ξ_{min}	-26 dB
N_e	.05 f_s

7.5.2 Selection of α

The accuracy of the *a priori* SIR (ξ) estimation plays a vital role in the performance of the spectral filters. Hence, we investigate the impact of the smoothing factor α on speech quality based on 25 random speech signals. Figure 7.3 shows PESQ and SRMR values for the competing methods under the assumption of an oracle PSD knowledge and noisy PSD estimation (added white noise to oracle PSD at 10 dB SNR). We observe from the plots that PESQ and SRMR show opposite behaviour as α increases. This indicates that a large value of α achieves better dereverberation at the cost of signal quality. It is also evident from the plot that GSS is the most sensitive to PSD estimation error which agrees our theoretical discussion in Section 7.3.4.

SS remains invariant to the change of α as it does not utilise ξ in its transfer function. WF follows the similar trend as GSS, however, GSS performs better with a better PSD estimation due to the inclusion of the corrective term γ. Among the other methods, STSA shows improved performance with an increasing α, however suffers degradation at $\alpha = 0.98$ in noisy PSD estimation. For the simulations in Section 7.5.3 and 7.5.4, we chose $\alpha = 0.9$ for STSA, $\alpha = 0.8$ for LSA, and $\alpha = 0.4$ for the remaining methods such that each individual method exhibits the best PESQ in Fig. 7.3 for the corresponding α. It is advisable to repeat this exercise



(a) PESQ - Oracle PSD

(b) PESQ - Noisy PSD

(c) SRMR - Oracle PSD

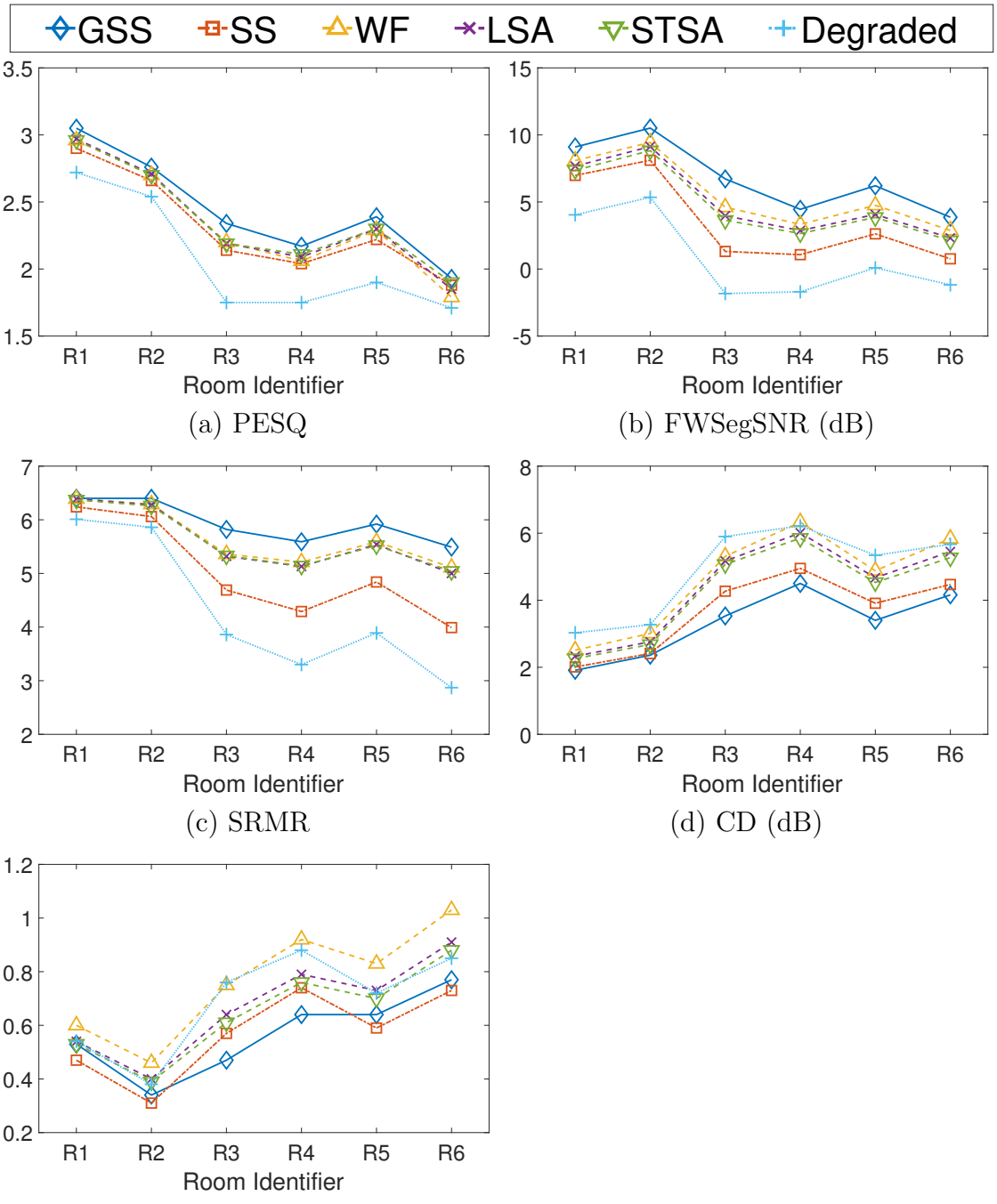
(d) SRMR - Noisy PSD

Figure 7.3: Performance of GSS with other conventional methods in terms of PESQ and SRMR for different α values with oracle PSD and noisy PSD estimation.

for each of the target audio scene in order to find the optimal value of α for the corresponding environment. Note that, we performed similar studies for η_x and β which did not exhibit any significant shift in the performance.

7.5.3 Performance based on oracle PSD knowledge

In the first comparative analysis, we included STSA [104] and LSA [105] with WF, SS and GSS. In this section, we considered oracle PSD knowledge to determine the true improvement of the competing techniques without the impact of PSD estimation accuracy. We used 25 random speeches in each reverberant condition



(a) PESQ

(b) FWSegSNR (dB)

(c) SRMR

(d) CD (dB)

(d) LLR

Figure 7.4: Comparison of GSS with other conventional approaches for oracle PSD knowledge.

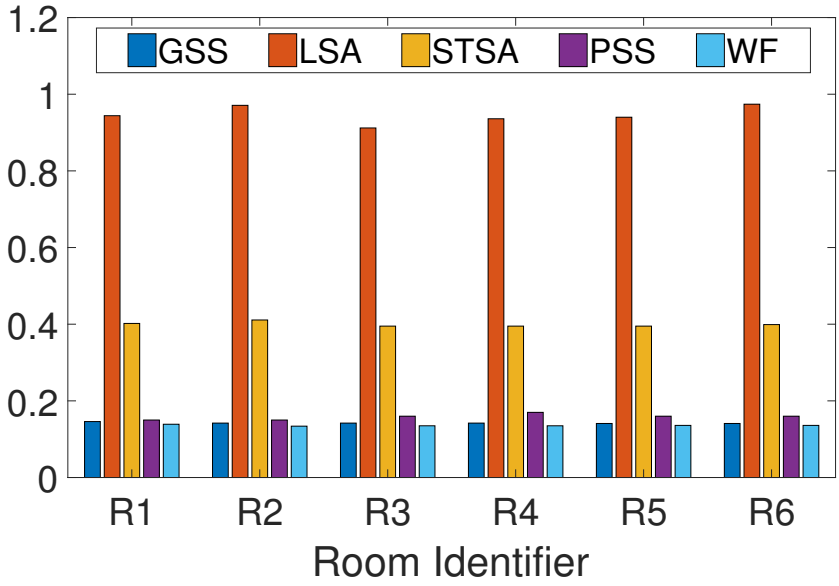


Figure 7.5: Average processing time per signal (assuming oracle PSD knowledge with 6.7s average signal duration).

with recorded air-condition noise at 10 dB SNR. The oracle PSDs were computed using the following exponential averaging formula

$$\Phi_r(\tau, k) = \eta\ \Phi_r(\tau - 1, k) + (1 - \eta)\ |X_r(\tau, k)|^2 \tag{7.27}$$

$$\Phi_n(\tau, k) = \eta\ \Phi_n(\tau - 1, k) + (1 - \eta)\ |X_n(\tau, k)|^2 \tag{7.28}$$

where $\eta = 0.85$ was used as the smoothing factor.

The results are shown in Fig. 7.4. It indicates that GSS consistently outperforms all other conventional methods, which is expected due to an accurate PSD estimation. The rest of the methods show a similar performance trend in terms of PESQ which is known to correlate well with overall quality. SS performs poorly in terms of suppressing reverberation, as evident from SRMR plot, but also manifest less signal distortion as evident from a low CD and LLR. The performance of WF, STSA, and LSA remain similar for most of the cases. Hence, it is fair to conclude that when the PSD estimation accuracy is high, GSS is the best selection among the lot.

In terms of processing time, SS, GSS, and WF were on the same level whereas

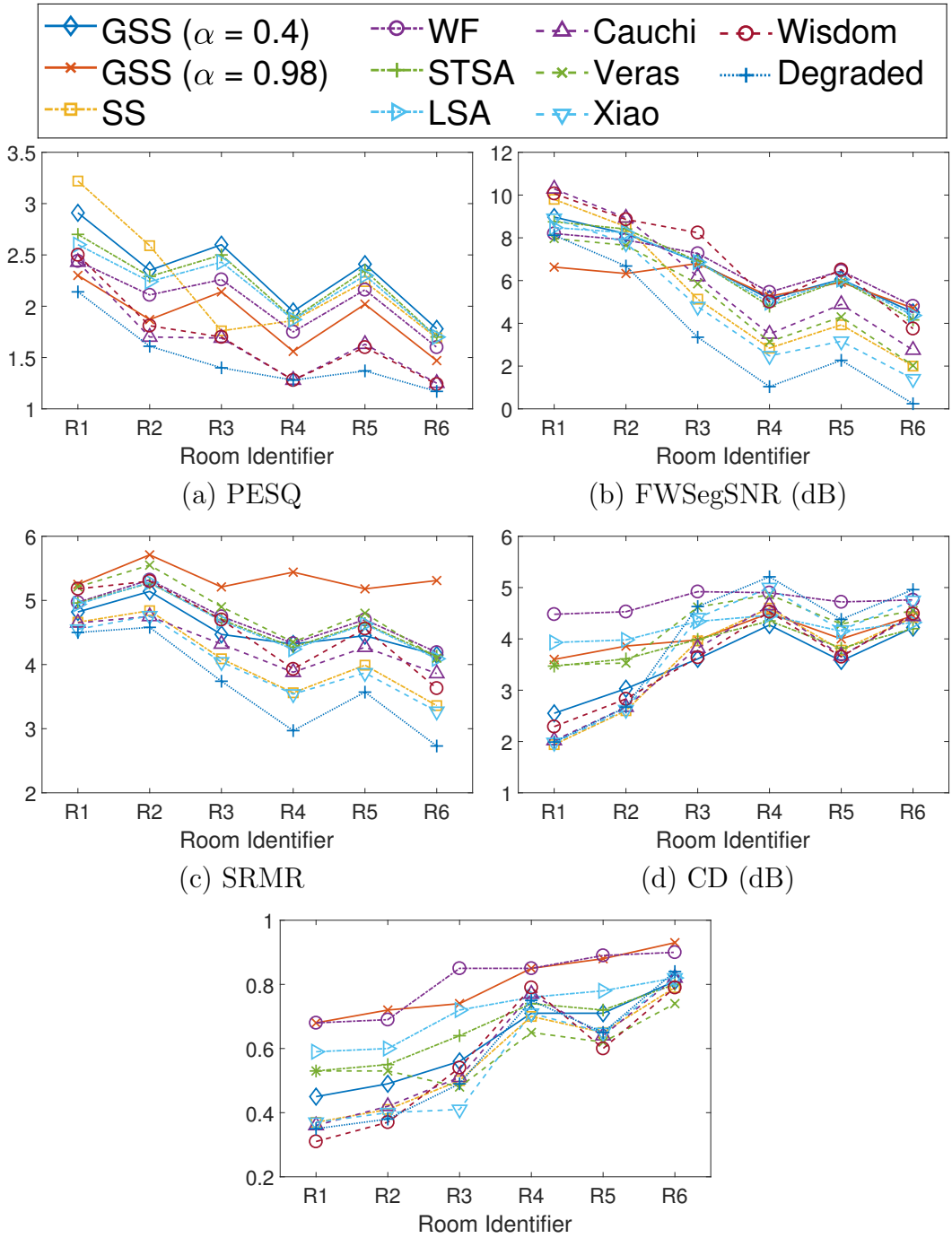
STSA and LSA took longer due to the requirements of extra Bessel and exponential function blocks. The relevant processing time is shown in Fig. 7.5.

7.5.4 Performance comparison based on estimated PSD

In this section, we analyse the performance of the aforementioned techniques using the estimated PSDs. We also compare them with the contemporary single-channel SS-based methods from the RCSE2014. We rigorously followed the instructions and procedures of RCSE2014 to offer a true comparison with the RCSE2014 methods. Note that, while we used the same PSD estimator for GSS and other conventional methods, the results of the RCSE2014 methods were directly fetched from the official source [214]. Hence, the accuracy of the PSD estimation need not be the same between the RCSE2014 and non-RCSE2014 methods.

Fig. 7.6 shows the comparative performances of GSS, 4 conventional methods, and 4 SS-based RCSE2014 methods [220]–[223]. The RCSE2014 methods in Fig. 7.6 are denoted by the name of the corresponding first author. As we updated the noise PSD only during the speech pauses, we used a lower smoothing constant $\eta_v = 0.6$ to put an extra confidence on the current frame estimate. However, we also applied a spectral floor of 0.2 in the unvoiced frames to compensate any rapid fluctuation in the noise PSD estimation. In the voiced frames where the *a priori* SIR is expected to be high, we used a spectral floor of 0.5 which resulted in a positive impact on the FWSegSNR.

We opted to investigate the advantage offered by GSS, if any, by evaluating it with two values of $\alpha = 0.4$ and 0.98 which will subsequently be identified as GSS_{40} and GSS_{98}, respectively. GSS_{40} exhibits relatively better quality under stronger reverberation. Increasing α to 0.98 results in a better reverberation suppression with GSS, however, degrades significantly in terms of other measures. WF performs better in terms of FWSegSNR, but shows signs of signal distortion based on CD and LLR. Conversely, SS shows notable improvement in room R1 and R2, but the performance quickly deteriorates as the reverberation gets stronger. This can be caused by the stronger cross-terms due to the correlation between the instant signal and reverberation spectra. STSA and LSA show mixed results, with STSA narrowly outperforms LSA in most of the cases.



(a) PESQ

(b) FWSegSNR (dB)

(c) SRMR

(d) CD (dB)

(d) LLR

Figure 7.6: Performance comparison using RCSE2014 dataset. PESQ was not reported for all the RCSE2014 methods.

In comparison with the RCSE2014 methods, GSS_{40} consistently performs better in terms of PESQ and CD in highly reverberant conditions, whereas maintaining a comparable performance in terms of other metrics. Hence, from the overall performance analysis, we can conclude that SS yields better performance in smaller rooms with weaker reverberation whereas GSS_{40} works better when reverberation is strong. Furthermore, we can observe that the selection of α directly influences the amount of dereverberation and speech distortion, and usually works in the opposite direction. With a smaller α, speech distortion remains at a lower level (i.e. higher PESQ value) at the cost of a less amount of dereverberation (i.e. low SRMR). Hence, α can be controlled to determine the trade off between dereverberation and speech distortion.

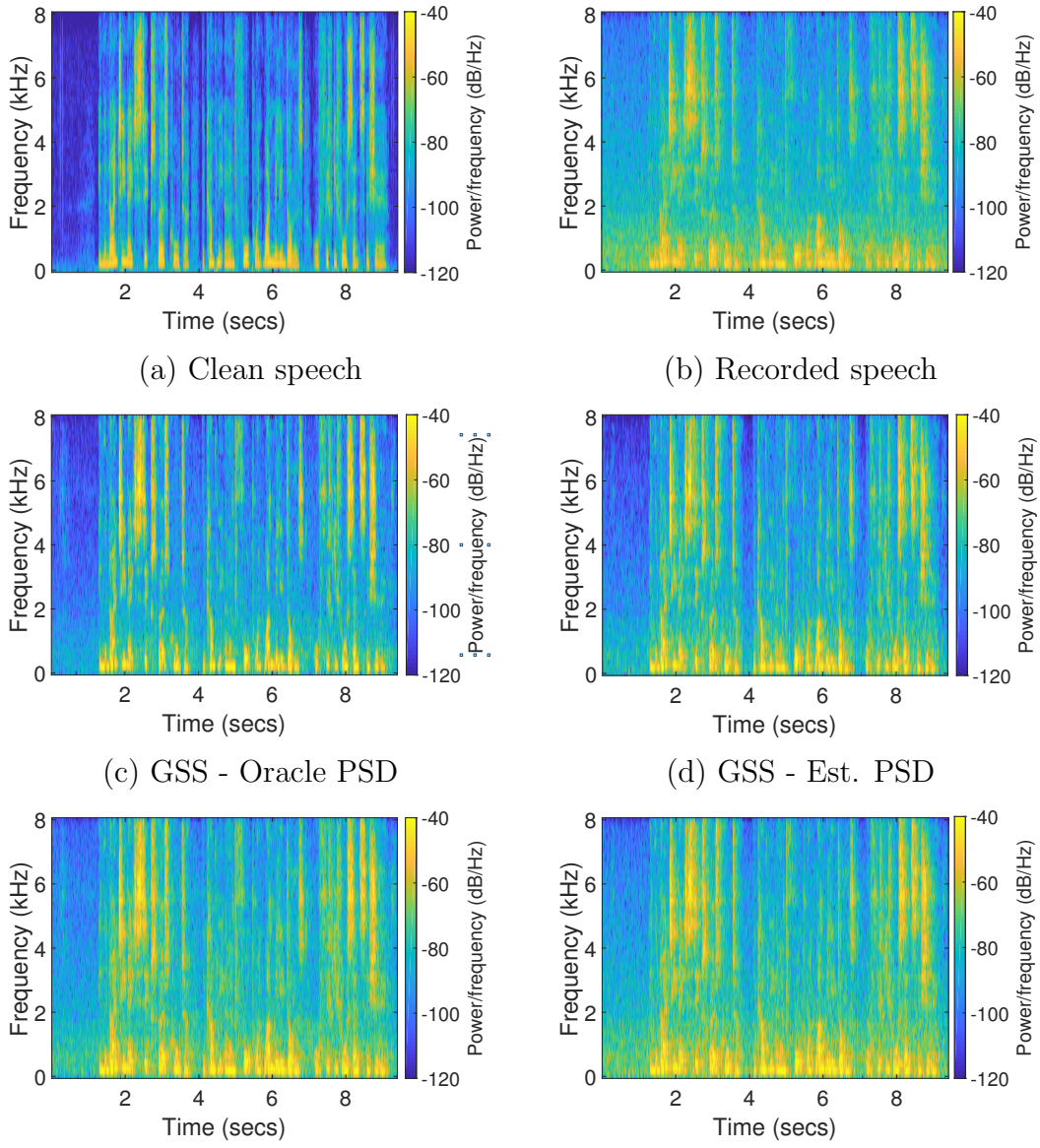
The performance issue of GSS in a less reverberant condition can be explained from Table 7.1 where we observe that the cross-term error is significantly lower at R1 and R2 ($T_{60} = 300$) compared to other room conditions ($T_{60} = 600, 700$). We discussed in Section 7.3.4 that the relative improvement of GSS is determined by the balance between cross-term improvement and *a priori* SIR estimation error. Hence, in a low reverberant environment where cross-terms remain low, *a priori* estimation error prevails and GSS shows an inferior performance.

For reference, Fig. 7.7 shows spectrograms of the processed speech signals for SS and GSS-based system outputs in room R5. It also shows that GSS outputs a cleaner spectrogram in a strong reverberant environment.

7.6 Summary

We conducted a case study with various post-filtering techniques to get an insight on their relative strengths and limitations. A theoretical analysis and detailed experimental validation are performed with a speech enhancement system in realistic noisy and reverberant environments. We discussed the fundamental limitations of WF, SS, and GSS, and explained the relation between PSD estimation accuracy and speech enhancement performance.

Through a series of experiments, we can conclude that while GSS improves the performance of conventional SS, it shows high sensitivity to PSD estimation error. Furthermore, it display better performance in strong reverberant environ-



(a) Clean speech

(b) Recorded speech

(c) GSS - Oracle PSD

(d) GSS - Est. PSD

(e) SS - Oracle PSD

(f) SS - Est. PSD

Figure 7.7: Spectrogram of the clean, degraded, and processed versions of a sample audio.

ment where the cross-term error is relatively high. Conversely, SS and WF perform better in a moderately reverberant environment.

7.7 Related Publication

- **A. Fahim**, P. N. Samarasinghe, and T. D. Abhayapala, "Single-Channel Speech Dereverberation in Noisy Environment for Non-Orthogonal Signals". Acta Acustica united with Acustica, Volume 104, Issue 6, pp. 1041–1055, 2018.

This page intentionally left blank.

Chapter 8

Conclusion and Future Work

In this chapter, we describe the major contributions made in this thesis. We also provide a brief overview on future research directions that can be derived from this work.

8.1 Conclusion

At the beginning of this thesis, we set our objective to exploit the spatial projection of a mixed soundfield to accomplish an efficient dissection to its primary components with respect to the source distribution. We aimed for two specific perceivable outcomes in terms of zonal separation of a soundfield and individual source separation. Both kinds of separation are extremely desirable in a large number of acoustic signal processing fields. The existing research on soundfield separation has been limited to planar separation whereas we aimed to formulate the solution for a more generic case. Conversely, source separation has seen a considerable amount of research pursuit resulting in the development of diverse algorithms, however, they are constrained by certain assumptions to deal with the challenges of non-stationary and dynamic nature of an audio signal. Hence, both the topics remain an active research area and the researchers are continuously seeking for improvements to address the existing constraints.

For zonal separation of a soundfield, we developed a novel technique that exploits the fundamentals of near-field acoustical holography with an array of higher

order microphones to achieve efficient and robust extraction of desired and undesired soundfields. We first validated the concept of soundfield separation using a height-invariant assumption of the acoustic environment. Later, we modified the model for higher order microphones in order to achieve a viable solution for a 3D soundfield. We analysed different practical aspects of both the models in order to gain an insight on their strengths and limitations. Unlike the planar separation which can only dissect a soundfield into two halves, the proposed algorithm is capable of estimating interior and exterior soundfields based on a finite bounded region. This concept facilitates compelling solutions to various acoustic problems such as selective soundfield recording for reproduction and active noise cancellation.

We then moved to accomplish individual source separation in a noisy and reverberant environments. Although, the soundfield separation technique can be used to complement a source separation algorithm by suppressing the undesired soundfield, we sought for a solution to source separation that can act independently in a noisy and reverberant environment. To this end, we first developed a mathematical model for the modal coherence of the spherical harmonic coefficients of a soundfield. We separately derived closed-form expressions for modal coherence of reverberant and a noise fields and consolidated them into a generic model for noisy and reverberant environment. The modal coherence model of a soundfield helped us to understand the soundfield behaviour and extract certain soundfield characteristics such as power spectral densities (PSD) of the soundfield components. It also allowed us to project soundfield behaviour based on the unique directional patterns of the modal coherence. We then trained a convolutional neural network to learn and predict source directions of arrival using a pattern recognition algorithm.

We demonstrated an application of the modal coherence-based PSD estimation in terms of sound source separation. In the first part of the demonstration, we exploited the full coherence matrix by estimating a complete set of spherical harmonic coefficients using a commercially available spherical microphone array. Later, we proposed a simpler planar array to achieve the source separation using a partial decomposition of a soundfield. The planar array provided an alternative approach to modal coherence-based separation with a reduced computational cost and a limited hardware support. We also carried out a detailed study on the selection of a post-filter for improving the interference rejection during the source

separation task. We analysed several existing approaches to spectral filtering and scrutinised their limitations and performance dependencies under various acoustic scenarios. The evaluation of source separation was performed using a real-world dataset which were measured in practical reverberant and noisy environments with a commercial microphone array. The exposure to the real environments and a commercial recording device allowed us to measure the robustness of the algorithms in the presence of various practical deviations such as background noise, microphone noise, and measurement inaccuracies.

The proposed algorithms for soundfield and source separation were found to be promising in solving several existing acoustical challenges. The evaluations revealed that these new algorithms can achieve better results in diverse acoustic environments, validated by the practical experiments. However, we identified certain areas that hold the potential of further research works to address a few existing challenges.

8.2 Future Work

Based on the foregoing discussion, we see potentials in the following research areas to improve the current state of the solution.

- **HOM array design:** One of the major challenges of extracting spherical harmonic coefficients is the required number of microphones which depends on the frequency range as well as the size of the target region. Hence, the practical usage of the proposed soundfield separation method is limited to specific applications, e.g., ANC, or needs to be constrained by additional priors. For a high fidelity speech processing over a large spatial zone, the 3D soundfield separation still requires a large number of microphones which can prove to be a hindrance in commercialising the concept. With the advent of novel harmonic extraction algorithms and rapid evolution in the technology, several compact and cost-effective alternatives for a higher order microphone have already been developed. However, further research is required in this field to improve the efficiency of harmonic extraction for a large-scale deployment of the proposed soundfield separation technique.

- **Real-time processing:** The modal coherence-based source separation exhibited better performances compared to the contemporary techniques. It was also found to be robust against different practical deviations and inaccuracies. However, the frequency-domain processing of the full coherence matrix can be resource intensive to perform in real-time. One of the solution is to employ the partial coherence matrix as we demonstrated with a planar microphone array. An alternative to this approach is to develop the modal correlation model using the real harmonics in the time domain to improve computational efficiency and latency of the system. The primary challenge to that approach is to extract the array-independent spherical harmonic coefficients $\alpha_{nm}(n)$ in the time domain. The existing time domain-based processing of spherical harmonics remained content with the array-dependent coefficients $\alpha_{nm}(n, r)$ due to the complexity in modelling the inverse Bessel function in the time domain. Active research is ongoing in this area to develop a closed-form expression of the time domain harmonics $\alpha_{nm}(n)$.

- **Robustness and resolution of DOA estimation:** We posed the DOA estimation as a classification task which inflicts a practical limitation on the spatial resolution of the algorithm. The resolution can be improved infinitely if the modal coherence patterns can be learned as a regression problem to predict the evolution between two trained classes. A second scope for improvement comes from the fact that CNN-based multi-source DOA estimation techniques generally assume W-disjoint orthogonality (WDO) which implies that each time-frequency (TF) bin in the STFT domain is dominated by a single source. However, as the number of sources increases, the rate of violation of WDO increases as well. In the original proposition of WDO [16], it was shown that in a 7-source mixture, approximately 72% TF bins obey the W-disjoint orthogonality. From the learning point of view, this does not incur a major issue as long as we can exclude the remaining TF bins from the processing. In our DOA estimation proposal, we used a probability-based filter to eliminate the ineffectual TF bins. However, CNN possess a tendency to best-match an unknown snapshot with an available class. Hence, the probability-based filtering might not have a full-proof solution in detect-

ing TF bins that breach WDO. This is a well-known open issue in machine learning domain, known as novelty detection, which is typically tried to solve with techniques like *one class SVM* or *open set recognition*. Hence, a detailed analysis is recommended in this topic to increase the robustness of the algorithm against WDO violation.

- **Tracking of moving sound sources:** The DOA estimation technique we developed in Chapter 6 was evaluated against multiple concurrent stationary sound sources. However, the method carries the potential to estimate DOA of moving sound sources due to its processing per time-frequency bin. Hence, the method inherently estimates DOA in each short time frames which makes it suitable to track fast-moving objects. Furthermore, the single-source training scheme for multi-source DOA estimation allows us to rely on a simple and efficient training stage even in a highly dynamic acoustic environment. To this end, it would require a fair amount of further research, tuning, and testing before this method can be successfully used in tracking moving sources which can be performed as an extension to this work.

This page intentionally left blank.

Appendix A

PSD estimation and source separation

A.1 The Definition of $W_{v,n,n'}^{u,m,m'}$

Defining

$$\overline{W}_{v,n,n'}^{u,m,m'} = \int_{\widehat{\boldsymbol{y}}} Y_{vu}(\widehat{\boldsymbol{y}})\ Y_{nm}(\widehat{\boldsymbol{y}})\ Y_{n'm'}(\widehat{\boldsymbol{y}})\ d\widehat{\boldsymbol{y}}, \tag{A.1}$$

the integral property of the spherical harmonics over a sphere suggests that

$$\overline{W}_{v,n,n'}^{u,m,m'} = \sqrt{\frac{(2v+1)(2n+1)(2n'+1)}{4\pi}} \begin{pmatrix} v & n & n' \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v & n & n' \\ u & m & m' \end{pmatrix} \tag{A.2}$$

where $(\cdot)$ in (A.2) represents Wigner-3j symbol [189]. Furthermore, a conjugated spherical harmonics coefficient follows the following property

$$Y_{nm}^{*}(\widehat{\boldsymbol{y}}) = (-1)^{m}\ Y_{n(-m)}(\widehat{\boldsymbol{y}}). \tag{A.3}$$

Henceforth, we use (A.1) and (A.3) to define

$$\begin{aligned} W_{v,n,n'}^{u,m,m'} &= \int_{\widehat{\boldsymbol{y}}} Y_{vu}(\widehat{\boldsymbol{y}})\ Y_{nm}^{*}(\widehat{\boldsymbol{y}})\ Y_{n'm'}(\widehat{\boldsymbol{y}})\ d\widehat{\boldsymbol{y}} \\ &= (-1)^{m} \int_{\widehat{\boldsymbol{y}}} Y_{vu}(\widehat{\boldsymbol{y}})\ Y_{n(-m)}(\widehat{\boldsymbol{y}})\ Y_{n'm'}(\widehat{\boldsymbol{y}})\ d\widehat{\boldsymbol{y}} \end{aligned}$$

$$= (-1)^m \ \overline{W}_{v,n,n'}^{u,-m,m'} \tag{A.4}$$

where the notation $W_{v,n,n'}^{u,m,m'}$ is chosen for brevity.

A.2 Closed-form expression of noise coherence matrix

Defining $\boldsymbol{x}'' = (\boldsymbol{x} - \boldsymbol{x}')$, where $\boldsymbol{x}'' \equiv (r'', \widehat{\boldsymbol{x}}'')$, we obtain (A.5) from the addition theorem for spherical Bessel functions in a similar manner as in [184, pp. 592-593]

$$Y_{a''b''}(\widehat{\boldsymbol{x}}'') j_{a''}(k \left\| \boldsymbol{x}'' \right\|) = 4\pi \sum_{ab}^{\infty} \sum_{a'b'}^{\infty} i^{(a-a'-a''+2b')} \ Y_{ab}(\widehat{\boldsymbol{x}}) \ j_a(k \left\| \boldsymbol{x} \right\|) \ Y_{a'b'}(\widehat{\boldsymbol{x}}') \ j_{a'}(k \left\| \boldsymbol{x}' \right\|) \ W_{a,a',a''}^{-b,b',b''}. \tag{A.5}$$

As $Y_{00}(\cdot) = 1/\sqrt{4\pi}$, we obtain (A.6) by letting $a'' = b'' = 0$ in (A.5)

$$j_0(k \left\| \boldsymbol{x}'' \right\|) = (4\pi)^{\frac{3}{2}} \sum_{ab}^{\infty} \sum_{a'b'}^{\infty} i^{(a-a'+2b')} \ Y_{ab}(\widehat{\boldsymbol{x}}) \ j_a(k \left\| \boldsymbol{x} \right\|) \ Y_{a'b'}(\widehat{\boldsymbol{x}}') \ j_{a'}(k \left\| \boldsymbol{x}' \right\|) \ W_{a,a',0}^{-b,b',0}. \tag{A.6}$$

Hence, using (A.6) in (4.35), we obtain

$$\Omega_{nm}^{n'm'}(k) = \frac{1}{|b_n(kr)|^2} \int_{\widehat{\boldsymbol{x}}} \int_{\widehat{\boldsymbol{x}}'} \Bigg((4\pi)^{\frac{3}{2}} \sum_{ab}^{\infty} \sum_{a'b'}^{\infty} i^{(a-a'+2b')} \times Y_{ab}(\widehat{\boldsymbol{x}}) \ j_a(k \left\| \boldsymbol{x} \right\|) \ Y_{a'b'}(\widehat{\boldsymbol{x}}') \ j_{a'}(k \left\| \boldsymbol{x}' \right\|) \ W_{a,a',0}^{-b,b',0} \Bigg) Y_{nm}^*(\widehat{\boldsymbol{x}}) \ Y_{n'm'}(\widehat{\boldsymbol{x}}') \ d\widehat{\boldsymbol{x}} \ d\widehat{\boldsymbol{x}}'. \tag{A.7}$$

Using the conjugate property of the spherical harmonics from (A.3) and rearranging (A.7), we obtain

$$\Omega_{nm}^{n'm'}(k) = \frac{(4\pi)^{\frac{3}{2}}}{|b_n(kr)|^2} \sum_{ab}^{\infty} \sum_{a'b'}^{\infty} i^{(a-a'+2b')} j_a(k \left\| \boldsymbol{x} \right\|) j_{a'}(k \left\| \boldsymbol{x}' \right\|)$$

$$W_{a,a',0}^{-b,b',0}\left(\int_{\widehat{\boldsymbol{x}}} Y_{ab}(\widehat{\boldsymbol{x}})\ Y_{nm}^{*}(\widehat{\boldsymbol{x}})\ d\widehat{\boldsymbol{x}}\right)\left((-1)^{m'}\int_{\widehat{\boldsymbol{x}}'} Y_{a'b'}(\widehat{\boldsymbol{x}}')\ Y_{n'(-m')}^{*}(\widehat{\boldsymbol{x}}')\ d\widehat{\boldsymbol{x}}'\right). \quad \text{(A.8)}$$

Finally, using they orthonormal property of the spherical harmonics from (2.34), we obtain

$$\Omega_{nm}^{n'm'}(k) = \frac{(4\pi)^{\frac{3}{2}} i^{(n-n')} j_n(kr) j_{n'}(kr)\ W_{n,n',0}^{-m,-m',0}}{|b_n(kr)|^2} \quad \text{(A.9)}$$

as $(-1)^{m'} = i^{2m'}$ and $\|\boldsymbol{x}\| = \|\boldsymbol{x}'\| = r$, where r is the radius of the spherical array.

A.3 Source directions

Table A.1 shows the true (θ, ϕ) and estimated $(\widehat{\theta}, \widehat{\phi})$ DOAs for $L = 4, 6$ based on a MUSIC-based algorithm run for individual sources.

Table A.1: Source directions in radian.

	Room A		Room B		Room C	
Source	θ, ϕ	$\widehat{\theta}, \widehat{\phi}$	θ, ϕ	$\widehat{\theta}, \widehat{\phi}$	θ, ϕ	$\widehat{\theta}, \widehat{\phi}$
4-speaker case						
S-01	$1.6, 5.81$	$1.6, 5.8$	$1.5, 0.75$	$1.5, 0.75$	$1.67, 0.77$	$1.64, 0.73$
S-02	$1.58, 4.53$	$1.58, 4.52$	$1.51, 2.31$	$1.51, 2.30$	$1.89, 2.33$	$1.93, 2.3$
S-03	$1.59, 3.19$	$1.59, 3.18$	$1.07, 4.01$	$1.13, 4.04$	$1.66, 3.87$	$1.7, 3.9$
S-04	$1.57, 1.93$	$1.57, 1.93$	$1.51, 5.4$	$1.52, 5.42$	$1.86, 5.44$	$1.89, 5.5$
6-speaker case						
S-01	$0.54, 5.56$	$0.52, 0.59$	$1.45, 6.23$	$1.46, 6.21$	$1.66, 6.23$	$1.64, 6.19$
S-02	$1.01, 3.51$	$1.02, 3.54$	$1.50, 0.74$	$1.51, 0.76$	$1.88, 1.01$	$1.89, 1.03$
S-03	$1.58, 5.17$	$1.55, 5.13$	$1.70, 1.46$	$1.72, 1.45$	$1.68, 2.04$	$1.69, 2.09$
S-04	$1.58, 1.32$	$1.55, 1.37$	$1.51, 2.31$	$1.52, 2.32$	$1.89, 3.11$	$1.83, 3.10$
S-05	$2.13, 2.85$	$2.15, 2.88$	$1.07, 4.02$	$1.09, 4.01$	$1.66, 4.25$	$1.62, 4.21$
S-06	$2.57, 6.18$	$2.56, 6.11$	$1.51, 5.41$	$1.52, 5.41$	$1.85, 5.19$	$1.84, 5.17$

This page intentionally left blank.

Bibliography

[1] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. London, UK: Academic press, 1999, vol. 108.

[2] J. D. Maynard, E. G. Williams, and Y. Lee, "Nearfield acoustic holography: I. Theory of generalized holography and the development of NAH," *The Journal of the Acoustical Society of America*, vol. 78, no. 4, pp. 1395–1413, 1985.

[3] J. F. Cardoso, "Eigenstructure of the 4th-order cumulant tensor with application to the blind source separation problem," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1989, pp. 2109–2112.

[4] P. Comon, "Independent component analysis, A new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[5] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[6] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[7] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[8] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, and Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1240–1250, 2013.

[9] B. D. Van Veen and K. M. Buckley, "Beamforming techniques for spatial filtering," *Digital Signal Processing Handbook*, pp. 61–1, 1997.

[10] J. Bourgeois and W. Minker, *Time-domain beamforming and blind source separation*. New York, USA: Springer-Verlag, 2010.

[11] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, 1998.

[12] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on antennas and propagation*, vol. AP-34, no. 3, pp. 276–280, 1986.

[13] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *Adaptive Antennas for Wireless Communications*, vol. 37, no. 7, pp. 984–995, 1989.

[14] J. H. Dibiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, 2001, pp. 157–180.

[15] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," PhD thesis, Brown University, 2000.

[16] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1846, 2004.

[17] P. Newell, *Recording studio design*. Routledge, 2012.

[18] J. G. Frayne, A. C. Blaney, G. R. Groves, and H. F. Olson, "A short history of motion-picture sound recording in the united states," *SMPTE Journal*, vol. 85, no. 7, pp. 515–528, 1976.

[19] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys*, vol. 48, no. 4, 2016.

[20] M. Cristani, M. Bicego, and V. Murino, "On-line adaptive background modelling for audio surveillance," in *International Conference on Pattern Recognition*, vol. 2, 2004, pp. 399–402.

[21] S. J. Elliott and P. A. Nelson, "Active noise control," *IEEE signal processing magazine*, vol. 10, no. 4, pp. 12–35, 1993.

[22] P. N. Samarasinghe, W. Zhang, and T. D. Abhayapala, "Recent advances in active noise control inside automobile cabins: Toward quieter cars," *IEEE Signal Processing Magazine*, vol. 33, no. 6, pp. 61–73, 2016.

[23] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets, "Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 31–42, 2015.

[24] K. Niwa, T. Nishino, and K. Takeda, "Encoding large array signals into a 3D sound field representation for selective listening point audio based on blind source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 181–184.

[25] E. A. Boniface, *Cockpit sound recorder*. US Patent 3,327,067, 1967. [Online]. Available: `https://www.google.com/patents/US3327067`.

[26] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[27] J. Wung, T. S. Wada, B. H. Juang, B. Lee, T. Kalker, and R. W. Schafer, "A system approach to residual echo suppression in robust hands-free teleconferencing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 445–448.

[28] S. Sakauchi, A. Nakagawa, Y. Haneda, and A. Kataoka, "Implementing and evaluating an audio teleconferencing terminal with noise and echo reduction," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2003, pp. 191–194.

[29] Y. J. Wu and T. D. Abhayapala, "Spatial multizone soundfield reproduction: Theory and design," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1711–1720, 2011.

[30] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 81–91, 2015.

[31] T. Betlehem and T. D. Abhayapala, "Theory and design of sound field reproduction in reverberant rooms," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2100–2111, 2006.

[32] J. Ahrens and S. Spors, "Sound field reproduction using planar and linear arrays of loudspeakers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2038–2050, 2010.

[33] M. M. Boone, E. N. G. Verheijen, and P. F. Van Tol, "Spatial sound-field reproduction by wave-field synthesis," *AES: Journal of the Audio Engineering Society*, vol. 43, no. 12, pp. 1003–1012, 1995.

[34] W. Jin and W. B. Kleijn, "Theory and design of multizone soundfield reproduction using sparse methods," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2343–2355, 2015.

[35] M. A. Poletti, T. D. Abhayapala, and P. N. Samarasinghe, "Interior and exterior sound field control using two dimensional higher-order variable-directivity sources," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 3814–3823, 2012.

[36] A. F. Metherell, H. M. El-Sum, J. J. Dreher, and L. Larmore, "Introduction to acoustical holography," *The Journal of the Acoustical Society of America*, vol. 42, no. 4, pp. 733–742, 1967.

[37] B. P. Hildebrand and B. B. Brenden, *Introduction to acoustical holography.* Plenum Press, NY, 1974.

[38] E. G. Williams and J. D. Maynard, "Holographic imaging without the wavelength resolution limit," *Physical Review Letters*, vol. 45, no. 7, pp. 554–557, 1980.

[39] M. Tamura, "Spatial Fourier transform method of measuring reflection coefficients at oblique incidence. I: Theory and numerical examples," *The Journal of the Acoustical Society of America*, vol. 88, no. 5, pp. 2259–2264, 1990.

[40] G. V. Frisk, A. V. Oppenheim, and D. R. Martinez, "A technique for measuring the planewave reflection coefficient of the ocean bottom," *The Journal of the Acoustical Society of America*, vol. 68, no. 2, pp. 602–612, 1980.

[41] E. Fernandez-Grande and F. Jacobsen, "Sound field separation with a double layer velocity transducer array (L)," *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 5–8, 2011.

[42] F. Jacobsen and Y. Liu, "Near field acoustic holography with particle velocity transducers," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3139–3144, 2005.

[43] E. Fernandez-Grande, F. Jacobsen, and Q. Leclère, "Sound field separation with sound pressure and particle velocity measurements," *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3818–3825, 2012.

[44] E. G. Williams, "Regularization methods for near-field acoustical holography," *The Journal of the Acoustical Society of America*, vol. 110, no. 4, pp. 1976–1988, 2001.

[45] R. Scholte, I. Lopez Arteaga, N. B. Roozen, and H. Nijmeijer, "Wavenumber domain regularization for near-field acoustic holography by means of modified filter functions and cut-off and slope iteration," *Acta Acustica united with Acustica*, vol. 94, no. 3, pp. 339–348, 2008.

[46] G. Chardon, L. Daudet, A. Peillot, F. Ollivier, N. Bertin, and R. Gribonval, "Near-field acoustic holography using sparse regularization and compressive sampling principles," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1521–1534, 2012.

[47] R. Steiner and J. Hald, "Near-field acoustical holography without the errors and limitations caused by the use of spatial DFT," *The International Journal of Acoustics and Vibration*, vol. 6, no. 2, pp. 83–89, 2001.

[48] F. Jacobsen and V. Jaud, "Statistically optimized near field acoustic holography using an array of pressure-velocity probes," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1550–1558, 2007.

[49] C.-X. Bi, X.-Z. Chen, and J. Chen, "Sound field separation technique based on equivalent source method and its application in nearfield acoustic holography," *The Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1472–1478, 2008.

[50] C.-X. Bi and J. Stuart Bolton, "An equivalent source technique for recovering the free sound field in a noisy environment," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 1260–1270, 2012.

[51] Y.-B. Zhang, F. Jacobsen, C.-X. Bi, and X.-Z. Chen, "Near field acoustic holography based on the equivalent source method and pressure-velocity transducers," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1257–1263, 2009.

[52] C.-X. Bi, D.-Y. Hu, Y.-B. Zhang, and J. S. Bolton, "Reconstruction of the free-field radiation from a vibrating structure based on measurements in a noisy environment," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 2823–2832, 2013.

[53] W. Q. Jia, J. Chen, C. Yang, and Z. Y. Wang, "Study of a sound field separation technique based on a single holographic surface and wave superposition method," *Journal of Mechanical Engineering Science*, vol. 223, no. 8, pp. 1827–1836, 2009.

[54] E. G. Williams and K. B. Washburn, "Generalized nearfield acoustical holography for cylindrical geometry: Theory and experiment," *The Journal of the Acoustical Society of America*, vol. 81, no. 2, pp. 389–407, 1987.

[55] M. Lee and J. S. Bolton, "Patch near-field acoustical holography in cylindrical geometry," *The Journal of the Acoustical Society of America*, vol. 118, no. 6, pp. 3721–3732, 2005.

[56] Y. T. Cho, J. S. Bolton, and J. Hald, "Source visualization by using statistically optimized near-field acoustical holography in cylindrical coordinates," *The Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2355–2364, 2005.

[57] N. P. Valdivia and E. G. Williams, "Study of the comparison of the methods of equivalent sources and boundary element methods for near-field acoustic holography," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 3694–3705, 2006.

[58] E. G. Williams, "The nearfield acoustical holography (NAH) experimental method applied to vibration and radiation in light and heavy fluids," *Computers and Structures*, vol. 65, no. 3, pp. 323–335, 1997.

[59] F. Jacobsen, G. Moreno-Pescador, E. Fernandez-Grande, and J. Hald, "Near field acoustic holography with microphones on a rigid sphere (L)," *The Journal of the Acoustical Society of America*, vol. 129, no. 6, pp. 3461–3464, 2011.

[60] H. M. Jones, R. A. Kennedy, and T. D. Abhayapala, "On dimensionality of multipath fields: Spatial extent and richness," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2002, pp. 2837–2840.

[61] P. N. Samarasinghe, T. D. Abhayapala, and M. A. Poletti, "Wavefield analysis over large areas using distributed higher order microphones," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 647–658, 2014.

[62] ——, "3D spatial soundfield recording over large regions," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[63] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2002, pp. 1781–1784.

[64] J. Daniel, "Evolving views on HOA: From technological to pragmatic concerns," in *Ambisonics Symposium*, 2009, pp. 1–18.

[65] H. Chen, T. D. Abhayapala, and W. Zhang, "Theory and design of compact hybrid microphone arrays on two-dimensional planes for three-dimensional soundfield analysis," *The Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 3081–3092, 2015.

[66] P. N. Samarasinghe, H. Chen, A. Fahim, and T. D. Abhayapala, "Performance analysis of a planar microphone array for three dimensional soundfield analysis," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 249–253.

[67] Z. Li and R. Duraiswami, "Flexible and optimal design of spherical microphone arrays for beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 702–714, 2007.

[68] T. D. Abhayapala and A. Gupta, "Spherical harmonic analysis of wavefields using multiple circular sensor arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1655–1666, 2010.

[69] E. De Sena, H. Hacihabiboglu, and Z. Cvetkovic, "On the design and implementation of higher order differential microphones," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 162–174, 2012.

[70] T. D. Abhayapala and A. Gupta, "Higher order differential-integral microphone arrays," *The Journal of the Acoustical Society of America*, vol. 127, no. 5, EL227–EL233, 2010.

[71] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2002, pp. 1949–1952.

[72] T. D. Abhayapala, R. A. Kennedy, and R. C. Williamson, "Nearfield broadband array design using a radially invariant modal expansion," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 392–403, 2002.

[73] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8, E1–E3.

[74] S. Brown and D. Sen, "Error analysis of spherical harmonic soundfield representations in terms of truncation and aliasing errors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 360–364.

[75] S. Amari, "A new learning algorithm for blind signal separation," *Advances in Neural Information Processing Systems*, vol. 8, pp. 757–763, 1996.

[76] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[77] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[78] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[79] M. A. Casey, "Separation of mixed audio sources by independent subspace analysis," in *International Computer Music Conference*, 2000, pp. 154–161.

[80] N. Mitianoudis and M. E. Davies, "Audio source separation of convolutive mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 489–497, 2003.

[81] I. Lee, T. Kim, and T. W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Processing*, vol. 87, no. 8, pp. 1859–1871, 2007.

[82] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 204–215, 2003.

[83] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[84] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Ninth International Conference on Spoken Language Processing*, vol. 5, 2006, pp. 2614–2617.

[85] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino, and S. Sagayama, "Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 5365–5368.

[86] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 1825–1828.

[87] D. Fitzgerald, M. Cranitch, and E. Coyle, "Shifted non-negative matrix factorisation for sound source separation," in *IEEE/SP 13th Workshop on Statistical Signal Processing*, 2005, pp. 1132–1137.

[88] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[89] Y. Wang and D. L. Wang, "Boosting classification based speech separation using temporal dynamics," *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, vol. 2, pp. 1526–1529, 2012.

[90] ——, "Cocktail party processing via structured prediction," *Advances in Neural Information Processing Systems*, vol. 1, pp. 224–232, 2012.

[91] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[92] X. Zhang and D. L. Wang, "Deep Learning Based Binaural Speech Separation in Reverberant Environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[93] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[94] Y. Shao and D. L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1589–1592, 2008.

[95] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[96] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[97] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," *International conference on latent variable analysis and signal separation*, pp. 258–266, 2017.

[98] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 3734–3738.

[99] G. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, no. 3-4, pp. 229–240, 1996.

[100] D. H. Johnson and D. E. Dudgeon, *Array signal processing: concepts and techniques*. NJ, USA: Prentice-Hall, Englewood Cliffs, 1993.

[101] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[102] W. Herbordt and W. Kellermann, "Adaptive Beamforming for Audio Signal Acquisition," in *Adaptive Signal Processing*, 2003, pp. 155–194.

[103] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[104] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, pp. 1109–1121, 1984.

[105] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, pp. 443–445, 1985.

[106] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1988, pp. 2578–2581.

[107] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374–2384, 2011.

[108] J. Benesty, J. Chen, Y. A. Huang, and S. Doclo, "Study of the Wiener filter for noise reduction," in *Speech Enhancement*, Berlin, Germany: Springer-Verlag, 2005, pp. 9–41.

[109] K. Lebart, J.-M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2014.

[110] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *European Signal Processing Conference (EUSIPCO)*, 2013, pp. 1–5.

[111] A. Kuklasinski, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1595–1608, 2016.

[112] H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Speech enhancement using nonlinear microphone array based on noise adaptive complementary beamforming," *IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Sciences*, vol. 83, no. 5, pp. 866–876, 2000.

[113] K. Niwa, T. Kawase, K. Kobayashi, and Y. Hioka, "PSD estimation in beamspace using property of M-matrix," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[114] J. Beh, D. Zotkin, and R. Duraiswami, "Adaptive interference rejection using generalized sidelobe canceller in spherical harmonics domain," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 47–51.

[115] L. Kumar and R. M. Hegde, "Near-field acoustic source localization and beamforming in spherical harmonics domain," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3351–3361, 2016.

[116] P. N. Samarasinghe, T. D. Abhayapala, and H. Chen, "Estimating the direct-to-reverberant energy ratio using a spherical harmonics-based spatial correlation model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 310–319, 2017.

[117] Y. Yamamoto and Y. Haneda, "Spherical microphone array post-filtering for reverberation suppression using isotropic beamformings," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[118] D. P. Jarrett, M. Taseska, E. A. P. Habets, and P. A. Naylor, "Noise reduction in the spherical harmonic domain using a tradeoff beamformer and narrowband DOA estimates," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 5, pp. 967–978, 2014.

[119] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Processing Magazine*, vol. 13, pp. 67–94, 1996.

[120] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 221–224.

[121] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 823–831, 1985.

[122] L. Birnie, T. D. Abhayapala, H. Chen, and P. N. Samarasinghe, "Sound source localization in a reverberant room using harmonic based MUSIC," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 651–655.

[123] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[124] A. Marti, M. Cobos, J. J. Lopez, and J. Escolano, "A steered response power iterative method for high-accuracy acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 2627–2630, 2013.

[125] L. O. Nunes, W. A. Martins, M. V. Lima, L. W. Biscainho, M. V. Costa, F. M. Gonçalves, A. Said, and B. Lee, "A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5171–5183, 2014.

[126] M. V. Lima, W. A. Martins, L. O. Nunes, L. W. Biscainho, T. N. Ferreira, M. V. Costa, and B. Lee, "A volumetric SRP with refinement step for sound source localization," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1098–1102, 2015.

[127] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction(SRC) on a large-

aperture microphone array," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2007, pp. 121–124.

[128] H. Ye and R. D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramer-Rao bounds for additive unknown colored noise," *IEEE Transactions on Signal Processing*, vol. 43, no. 4, pp. 938–949, 1995.

[129] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 1843–1854, 2002.

[130] P. Stoica and K. C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 7, pp. 1132–1143, 1990.

[131] M. I. Mandel, D. P. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Information Processing Systems*, 2007, pp. 953–960.

[132] O. Schwartz, Y. Dorfan, E. A. P. Habets, and S. Gannot, "Multi-speaker DOA estimation in reverberation conditions using expectation-maximization," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[133] O. Schwartz, Y. Dorfan, M. Taseska, E. A. P. Habets, and S. Gannot, "DOA estimation in noisy environment with unknown noise power using the EM algorithm," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2017, pp. 86–90.

[134] X. Li, L. Girin, R. Horaud, and S. Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1997–2012, 2017.

[135] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1913–1928, 2010.

[136] C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *The Journal of the Acoustical Society of America*, vol. 108, no. 4, pp. 1888–1905, 2000.

[137] B. Rafaely and D. Kolossa, "Speaker localization in reverberant rooms based on direct path dominance test statistics," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 6120–6124.

[138] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.

[139] D. Levin, E. A. P. Habets, and S. Gannot, "On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1800–1811, 2010.

[140] A. H. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *European Signal Processing Conference (EUSIPCO)*, 2015, pp. 2296–2300.

[141] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 178–192, 2017.

[142] S. Hafezi, A. H. Moore, and P. A. Naylor, "Augmented intensity vectors for direction of arrival estimation in the spherical harmonic domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1956–1968, 2017.

[143] S. Tervo and A. Politis, "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1539–1551, 2015.

[144] S. Hafezi, A. H. Moore, and P. A. Naylor, "3D acoustic source localization in the spherical harmonic domain based on optimized grid search," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2016-May, 2016, pp. 415–419.

[145] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 2814–2818.

[146] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6.

[147] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 8, pp. 6403–6413, 2018.

[148] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 2386–2390.

[149] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 405–409.

[150] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal on Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[151] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal on Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[152] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *European Signal Processing Conference (EUSIPCO)*, vol. 2018-Septe, 2018, pp. 1462–1466.

[153] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, no. 6, pp. 453–466, 2008.

[154] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, 2017.

[155] D. Menzies and M. Al-Akaidi, "Nearfield binaural synthesis and ambisonics," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1559–1563, 2007.

[156] B. Bernschutz, A. V. Giner, C. Porschmann, and J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[157] M. Noisternig, F. Zotter, and B. F. G. Katz, "Reconstructing sound source directivity in virtual acoustic environments," in *Principles and Applications of Spatial Hearing*, 2011, pp. 357–372.

[158] P. N. Samarasinghe, T. D. Abhayapala, M. A. Poletti, and T. Betlehem, "An efficient parameterization of the room transfer function," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2217–2227, 2015.

[159] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2828–2840, 2012.

[160] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 261–270, 2012.

[161] E. Tiana-Roig, F. Jacobsen, and E. Fernandez-Grande, "Beamforming with a circular array of microphones mounted on a rigid sphere (L)," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1095–1098, 2011.

[162] S. Delikaris-Manias, J. Vilkamo, and V. Pulkki, "Signal-dependent spatial filtering based on weighted-orthogonal beamformers in the spherical harmonic domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1507–1519, 2016.

[163] C. C. Lai, S. Nordholm, and Y. H. Leung, "Design of steerable spherical broadband beamformers with flexible sensor configurations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 427–438, 2013.

[164] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 361–371, 2011.

[165] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the Audio Engineering Society*, vol. 53, no. January, pp. 1004–1025, 2016.

[166] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on speech and audio processing*, vol. 9, no. 6, pp. 697–707, 2001.

[167] W. Zhang, P. N. Samarasinghe, H. Chen, and T. D. Abhayapala, "Surround by sound: A review of spatial audio recording and reproduction," *Applied Sciences*, vol. 7, no. 5, 2017.

[168] B. Bu, C.-c. Bao, and M. S. Jia, "Design of a planar first-order loudspeaker array for global active noise control," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2240–2250, 2018.

[169] S. Spors and H. Buchner, "Efficient massive multichannel active noise control using wave-domain adaptive filtering," *International Symposium on Communications, Control, and Signal Processing (ISCCSP)*, pp. 1480–1485, 2008.

[170] J. Zhang, T. D. Abhayapala, W. Zhang, P. N. Samarasinghe, and S. Jiang, "Active noise control over space: A wave domain approach," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 774–786, 2018.

[171] F. Ma, W. Zhang, and T. D. Abhayapala, "Active control of outgoing noise fields in rooms," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1589–1599, 2018.

[172] A. H. Moore and P. A. Naylor, "Linear prediction based dereverberation for spherical microphone arrays," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.

[173] Y. Peled and B. Rafaely, "Method for dereverberation and noise reduction using spherical microphone arrays," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 113–116.

[174] D. P. Jarrett, E. A. P. Habets, M. R. Thomas, N. D. Gaubitch, and P. A. Naylor, "Dereverberation performance of rigid and open spherical microphone arrays: Theory & simulation," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011, pp. 145–150.

[175] M. Caresta and N. J. Kessissoglou, "Acoustic signature of a submarine hull under harmonic excitation," *Applied Acoustics*, vol. 71, no. 1, pp. 17–31, 2010.

[176] S. Wang, "Finite-difference time-domain approach to underwater acoustic scattering problems," *The Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 1924–1931, 1996.

[177] U. Bolleter and M. J. Crocker, "Theory and measurement of modal spectra in hard-walled cylindrical ducts," *The Journal of the Acoustical Society of America*, vol. 51, no. 5A, pp. 1439–1447, 1972.

[178] R. P. Radlinski and T. J. Meyers, "Radiation patterns and radiation impedances of a pulsating cylinder surrounded by a circular cage of parallel cylindrical rods," *The Journal of the Acoustical Society of America*, vol. 56, no. 3, pp. 842–848, 1974.

[179] Y. J. Wu and T. D. Abhayapala, "Theory and design of soundfield reproduction using continuous loudspeaker concept," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 107–116, 2009.

[180] M. A. Poletti, T. Betlehem, and T. D. Abhayapala, "Higher-order loudspeakers and active compensation for improved 2d sound field reproduction in rooms," *AES: Journal of the Audio Engineering Society*, vol. 63, no. 1-2, pp. 31–45, 2015.

[181] M. R. Thomas, J. Ahrens, and I. Tashev, "A method for converting between cylindrical and spherical harmonic representations of sound fields," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 4723–4727.

[182] Z. Li, R. Duraiswami, and N. A. Gumerov, "Capture and recreation of higher order 3D sound fields via reciprocity," in *International Conference on Auditory Display (ICAD)*, 2004.

[183] R. Rabenstein, S. Spors, and P. Steffen, "Wave field synthesis techniques for spatial sound reproduction," *Topics in Acoustic Echo and Noise Control*, vol. 5, pp. 517–545, 2006.

[184] W. C. Chew, *Waves and fields in inhomogeneous media*. IEEE press New York, 1995, vol. 522.

[185] D. Colton and R. Kress, *Inverse acoustic and electromagnetic scattering theory*, 3rd ed. Berlin, Germany: Springer-Verlag, 2012, vol. 93.

[186] H. Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*. Berlin, Germany: Springer, 2007, vol. 348.

[187] P. N. Samarasinghe, "Modal based solutions for the acquisition and rendering of large spatial soundfields," PhD thesis, Australian National University, 2015.

[188] A. R. Edmonds, *Angular momentum in quantum mechanics*. Princeton, New Jersey: Princeton university press, 1957.

[189] F. W. J. Olver, "3j, 6j, 9j symbols," in *NIST handbook of mathematical functions*, Cambridge, UK: Cambridge University Press, 2010, ch. 34, pp. 755–766.

[190] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[191] P. A. Martin, *Multiple scattering: interaction of time-harmonic waves with N obstacles*, 2006.

[192] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: a British English speech corpus for large vocabulary continuous speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 81–84.

[193] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[194] F. Borra, I. D. Gebru, and D. Markovic, "Soundfield Reconstruction in Reverberant Environments Using Higher-order Microphones and Impulse Response Measurements," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2019-May, 2019, pp. 281–285.

[195] H. Zuo, P. N. Samarasinghe, and T. D. Abhayapala, "Exterior-interior 3D sound field separation using a planar array of differential microphones," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 216–220.

[196] P. A. Naylor and N. D. Gaubitch, "Introduction," in *Speech dereverberation*, London, UK, 2010, pp. 1–15.

[197] P. D. Teal, T. D. Abhayapala, and R. A. Kennedy, "Spatial correlation for general distributions of scatterers," *IEEE Signal Processing Letters*, vol. 9, no. 10, pp. 305–308, 2002.

[198] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[199] MH Acoustics, "EM32 Eigenmike microphone array release notes (v17. 0)," *25 Summit Ave, Summit, NJ 07901, USA*, 2013. [Online]. Available: `https://mhacoustics.com/`.

[200] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge - corpus description and performance evaluation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015, pp. 1–5.

[201] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," *Linguistic data consortium*, 1993.

[202] N. Moritz, M. R. Schadler, K. Adiloglu, B. T. Meyer, T. Jurgens, T. Gerkmann, B. Kollmeier, S. Doclo, and S. Goetze, "Noise robust distant automatic speech recognition utilizing NMF based source separation and auditory feature extraction," in *Proc. of CHiME*, 2003, pp. 1–6.

[203] S. Yamamoto, J. M. Valin, K. Nakadai, J. Rouat, F. Michaud, T. Ogata, and H. G. Okuno, "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," in *IEEE International Conference on Robotics and Automation*, vol. 2005, 2005, pp. 1477–1482.

[204] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[205] B. Rafaely and M. Kleider, "Spherical microphone array beam steering using Wigner-D weighting," *IEEE Signal Processing Letters*, vol. 15, pp. 417–420, 2008.

[206] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised source localization on multiple manifolds with distributed microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1477–1491, 2017.

[207] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *International Workshop on Acoustic Echo and Noise Control*, 2008, pp. 751–758.

[208] S. Hafezi, A. H. Moore, and P. A. Naylor, "Robust source counting and acoustic DOA estimation using density-based clustering," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2018, pp. 395–399.

[209] E. A. P. Habets, *Room impulse response generator*, 2006. [Online]. Available: `https://github.com/ehabets/RIR-Generator`.

[210] J. Fliege and U. Maier, "A two-stage approach for computing cubature formulae for the sphere," in *Mathematik 139T, Universitat Dortmund, Fachbereich Mathematik, Universitat Dortmund, 44221*, 1996, pp. 1–31.

[211] F. Chollet *et al.*, *Keras*, 2015. [Online]. Available: `https://keras.io`.

[212] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, D. Andy, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous distributed systems*, 2015. [Online]. Available: `https://www.tensorflow.org/`.

[213] E. A. Robinson and S. Treitel, "Principles of Digital Wiener Filtering," *Geophysical Prospecting*, vol. 15, no. 3, pp. 311–332, 1967.

[214] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *Eurasip Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, 2016.

[215] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[216] S. V. Vaseghi, *Advanced digital signal processing and noise reduction*. John Wiley & Sons: John Wiley & Sons, 2008, vol. 9.

[217] J. Polack, "La transmission de l'énergie sonore dans les salles," PhD thesis, Le Mans, 1988.

[218] H. W. Löllman, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *International Workshop on Acoustic Echo and Noise Control*, 2010, pp. 1–4.

[219] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[220] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukic, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme," in *REVERB Challenge Workshop*, 2014, pp. 1–8.

[221] J. C. S. Veras, T. d. M. Prego, A. A. de Lima, T. N. Ferreira, and S. L. Netto, "Speech quality enhancement based on spectral subtraction," in *REVERB Challenge Workshop*, 2014.

[222] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "The NTU-ADSC systems for reverberation challenge 2014," in *REVERB Challenge Workshop*, o2.2, 2014.

[223] S. Wisdom, T. Powers, L. Atlas, and J. Pitton, "Enhancement of reverberant and noisy speech by extending its coherence," in *REVERB Challenge Workshop*, 2014. [Online]. Available: `http://arxiv.org/abs/1509.00533`.